DATASCI 1000

Final Exam
Booklet Solutions
Winter 2026

DATASCI 1000 Final Exam Booklet Solutions (Winter 2026)

A. Variables

Describe the Shape of Each Stemplot below:

a) This is a symmetrical or bell-shaped distribution.
b) This is a left-skewed distribution as the long tail is to the left.
c) This is a uniform shaped distribution.

Example 1.

Stem	Leaves
3	4
4	5
5	
6	5 6 7
7	6 7
8	0 1 9
9	

Example 2.
A. is a graph used to graph one quantitative variable

Example 3.
B. a categorical variable

Example 4.

C. is correct since number of years and high school average are quantities while classes and gender are just categories

Example 5.

A. height is NOT a category, it is a quantity

A1. You should use a bar graph or a pie chart, since the data is categorical. The answer is b).

A2. Sex and name of university are categorical variables and the rest are quantitative. The answer is c).

A3. You must include all stems, even if there are no data in them
a) cannot be correct because 20 on the left of the line and 3 on the right, would mean 203 and not 23 as required.

The answer is b).

A4. The outlier is 11%. The answer is b).

A5. Colour is a categorical variable and therefore we can use a bar or a pie graph to display it. The answer is d).

A6.This data is measured in MPG and it is quantitative data. Therefore, the answer is c).

A7. c) doesn't show any data above 80 and the graph clearly does, so it is incorrect.
b) is incorrect because a stemplot cannot skip numbers in the stems, ie. 2,3,4,5,8,20 is missing numbers

So, the correct stemplot is a).

A8.No, there is only one "maximum" bar. This is false.

A9.Age, height, amount of student loans and present annual salary are quantitative variables. Present major and plans after graduation are categorical variables.

A10. Flip the graph on its side and it is right-skewed. The lowest mark is 52% and the highest mark is 85%. The answer is d).

A11. Categorical are: amenities, inclusive or not, location

Quantitative variables are: price per night, average room size and resort size

A12. B. is categorical since colour is not a quantity.

B. Measures of Central Tendency

Example 1.

C. is not a measure of central location; it is a measure of spread of the data.

Example 2.

The answer is D. if there are two numbers in the middle, we average them. Data must be in increasing order first.

Example 3.

Positively skewed is skewed to the right, so the mean is pulled towards the tail, to the right. So, the mean will be greater than the median and the answer is A.

Example 4.

4, 5, 5, 5, 6, 6, 7, 8, 11, 13

(a) The median is the middle # = average of 6 and 6 = 6...answer is B

(b) The mode is 5, since it occurs the most often...answer is A

(c) To find the mean, add up all data and divide by 10 numbers...mean is 7...answer is C.

B1. Mode- occurs most often...Therefore, 67 and 78 (bi-modal)

Median- Write numbers in ascending order and take the middle # which is 67.
34, 44, 50, 56, 66, 67, 67, 78, 78, 88, 98

Mean=$\frac{34+44+\cdots+98}{11}$ =66

B2.
Median= 70 (middle # when written in ascending order)

Mean=$\frac{60+70+80}{3} = 70$

After adding a mark of 75... the numbers are 60,70,75,80

Median=$\frac{70+75}{2} = 72.5$

Mean=$\frac{60+70+75+80}{4} = 71.25$

The answer is c).

B3.

$$\frac{80 + 75 + 95 + x}{4} = 80.5$$

80+75+95+x = 322

x = 322 - 95 - 75 - 80
x=72

Therefore, the mark on the fourth test was 72.

B4. The numbers represented by the stemplot are 54, 56, 62, 63, 65, 68, 71, 74, 83, 92
Median=$\frac{65+68}{2} = 66.5$
Mean=$\frac{54+56+\cdots+92}{10} = 68.8$

B5. 64, 70, 74, 80, 92 the median is 74. The answer is a).

B6. The answer is (c).

B7. Both a stemplot and boxplot reveal the shape as if you flip them sideways, you can tell the skew from both of them. Stemplots are NOT better for large data sets, since they display every number across the page. So, the answer is III. since a stemplot does show every number. the answer is C.

B8. n=4+7+3+3+2+1=20 data
notice this time the y-axis isn't percent, but is frequency or number
median occurs at (n+1)/2 = 21/2 = 10.5...median is the average of the 10th and 11th data

add heights of bars...first two would be 4+7 = 11th data, so the 10th and 11th occur in this bar and the median is between $1.00 and $1.50.

B9. This has a tail to the right, so it is positively or right skewed. The answer is C.

B10. Since there are 119 students, the median occurs at (n+1)/2=(119+1)/2=60th data...

The first four bars are 1+2+15+24=42%
0.42x119=50th data, but we need the 60th, so keep adding bars

1+2+15+24+31=73%
0.73x119=87>60th, so the median occurs in this bar.

Therefore, the median is approximately 10 pounds. The answer is B.

C. Measures of Spread

Example 1. Leave out the median=12 since there are an odd number of data

Q1=median of the bottom half= median of 10 and 10 = 10
Q3= median of the top half=median of 18 and 20 = 19

Therefore, the first quartile is 10 and the third quartile is 19.

Example 2.

The median is the middle number when the numbers are written in increasing order. There are 10 data, so it is the average of the 5th and 6th pieces of data. Median=(12+15)/2=13.5

Put the numbers in order and then since there are 10 numbers, the first quartile is the median of the bottom five numbers...Q1=10
Q3= median of the top five numbers=20

IQR=Q3-Q1=20-10=10

Example 3.

First, find the mean...mean=average of the numbers=$\frac{4+5+6+3+2}{5} = 4$
Variance=$\frac{(4-4)^2+(5-4)^2+(6-4)^2+(3-4)^2+(2-4)^2}{4} = \frac{0+1+4+1+4}{4} = 2.5$

Standard deviation=$\sqrt{V(x)} = \sqrt{2.5}$=1.58

Example 4.

$$\bar{x} = \frac{195}{5} = 39$$
$$s^2 = \frac{(35-39)^2+(25-39)^2+(75-39)^2+(15-39)^2+(45-39)^2}{4}$$
$$s^2 = \frac{16+196+1296+576+36}{4}$$
$$s^2 = 530$$
$$s = \sqrt{530}$$
$$= 23.02$$

<u>Example 5.</u>

6, 14, 67, 73, 73   74, 87, 90, 95, 99

a) 73.5

b) $\frac{678}{10} = 67.8$

c) $s = \sqrt{\frac{(6-67.8)^2+(14-67.8)^2+\cdots+(99-67.8)^2}{9}} = 32.29$

d) $Q1 = 3rd\ number = 67$
$Q3 = 8th\ number = 90$
$IQR = 90 - 67 = 23$

e) Below $Q1 - 1.5(IQR) = 67 - 1.5(23) = 32.5$ ∴ 6,14 are outliers
Above $Q3 + 1.5(IQR) = 90 + 1.5(23) = 124.5$ $\therefore none$

f) $min = 6$ $Q3 = 90$
$Q1 = 67$ $max = 99$
$Q2 = 73.5$

g)

Standard Boxplot

diff

```
|----|===================|===========|
6    14    67       73.5       90         99
```

(Shows all points as part of the range — doesn't mark outliers separately.)

Modified Boxplot (with outliers shown as dots)

diff

```
  o    o          |=====|=========|-----|
 6    14         67     73.5      90     99
```

- The **box** spans from Q1 (67) to Q3 (90).
- The **line** inside the box marks the median (73.5).
- The **whiskers** extend from 67 to 99.
- The **dots (o)** represent outliers at 6 and 14.

Example 6.

A). Yellow board...median is closest to 45...answer is (b).

B). Green board Q1=25 and Q3=33 and IQR=33-25=8
answer is (b).

C). The shape of the beetles on green boards is left skewed or negatively skewed. The answer is (a). (long tail to the left, it you flip the Box plot on its side)

Example 7.

(a) Blue IQR=10

White Q1=12 and Q3=20, so IQR=20-12=8, so blue has a larger IQR.

(b) Green maximum=37 and White maximum=23, so green is greater

(c) White board IQR=8

Outliers occur below Q1-1.5(IQR)= 12 - 1.5(8)=0 so below 0...none
or

Outliers occur above Q3+1.5(IQR)=20+1.5(8)=32 above 32...none

There are no outliers for the white board.

Example 8.

This graph is skewed to the right, so we need to use a measure that is resistant. The mean is not a resistant measure of the centre. The standard deviation is not resistant and since the question asks about centre, it couldn't be standard deviation anyway as it measures spread. So, we would use the median, since it is a measure of centre and it IS resistant.

The answer is (b).

Example 9.

a) median occurs at n+1/2 = 100+1/2 = 101/2=50.5...average of the 50th and 51st numbers

5+18= 23...not in the second bar

5+18+42=65...too far, so the median is in the third bar...between 66 and 69

b) first quartile= 25% lies below it...so 25 people below it...so it would be in the 66 to 69 range as well since 5+18=23 isn't quite 25%. The first quartile would be the average of the 25th and 26th numbers, so 66 to 69.

c) the third quartile means 75% lie below it or 75 people lie below it.
5+18+42+27=92...so it is somewhere after the third bar since it added up to 65 which wasn't large enough, so the third quartile is in between 69 and 72.

C1. a) is false because the standard deviation is NOT resistant.

C2. If all of the data are equal, the mean will be whatever that value is...for example for the sample: 5,5,5,5,5,5, the mean is 5. The variance and standard deviations will be 0. For the IQR, Q1=5 and Q3=5, so IQR=5-5=0
The answer is (d).

C3.a) This graph is skewed to the right, so the mean is pulled to the right tail. The mean is the largest and so the answer is ii).

b) He has received fewer than 50 spam emails 6+5+5=16 days out of 20 days=80%. The answer is iv).

c) check to see if there are outliers...like the number 105?
Q1=25 (average of 5^{th} and 6^{th} data=25+25/2)
Q3=48(average of 15^{th} and 16^{th} numbers = (47+49)/2=48)
IQR=Q3-Q1=48-25=23

outliers occur below Q1-1.5(IQR)= 25 – 1.5(23) = -9.5 none below this
above Q3+1.5(IQR)=48+1.5(23)=82.5 so the number 105 is above 82.5 and is an outlier

C4.
24, 46, 49, 51, 64, 64,** 67,** 81, 88, 89, 97, 103,120

If we look, we can see that the data is in order. There are 13 data points, so the median is the 7th data which is 67.

Q_1 is in between the 3rd and 4th data points, and Q_3 is between the 10th and 11th data points.
$Q_1 = 50$
median=67
$Q_3 = 93$
$Range = 120 - 24 = 96$
$IQR = 93 - 50 = 43$
We can use the formulas above to calculate the mean and sample standard deviation:

$$\bar{x} = \frac{24 + 46 + \cdots 97 + 103 + 120}{13} = 72.5$$

$$s = \sqrt{\frac{(24-72.5)^2 + (46-72.5)^2 + \cdots (97-72.5)^2 + (103-72.5)^2 + (120-72.5)^2}{12}} = \sqrt{\frac{8615.25}{12}} = 26.8$$

We now check for outliers:

$$Q_3 + 1.5\ IQR = 93 + 1.5(43) = 157.5$$

$$Q_1 - 1.5\ IQR = 50 - 1.5(43) = -14.5$$

Since no data point is above 157.5 or below -14.5, there are no outliers in the data set.

C5. **(a)**

Unimodal (one peak).
Asymmetric – skewed to the left.
Has a suspected outlier (155g).
With the wrong data: $\bar{x} = \frac{\sum x}{n}$
Since the original mean is 218, we can multiply by n=37 and find the sum of the 37 reactions, and we get: 218(37)=8066
The new mean involves subtracting the incorrect number and adding the correct one to the sum and then finding the new mean by dividing by 37 reactions

With the correct data: $\bar{x} = \frac{8066-155+195}{37} = \frac{8106}{37} = 219.1$

(ii) How will the values of the following summary statistics change after the data correction is made?

Choose your answer from the following list:

I. The value becomes smaller after the correction is made.
II. The value becomes larger after the correction is made.
III. There is no change in the value after the correction is made.

Summary Statistics	
median	III
IQR	III
standard deviation	I
75th percentile	III

C6.

Write the numbers in ascending order: n=9

35, 50, 55, 65, 65, 70, 80, 80, 95

Mode=65, 80 (bi-modal)

Median=middle # = 5th number = 65

Mean=$\frac{35+50+\cdots+95}{9} = 66.1 \cong 66$

Range= max - min= 95 - 35 = 60

Make a chart to find the standard deviation:

$x_i - \bar{x}$	$(x_i - \bar{x})^2$
35-66	961
50-66	256
55-66	121
65-66	1
65-66	1
70-66	16
80-66	196
80-66	196
95-66	841

$$s = \sqrt{\sum \frac{(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{2589}{8}} = 17.99$$

C7.

55, 56, 66, *76*, 77, 88,89

Min= 55
Q1= 56 (middle of bottom half)
Median= 76
Q3=88 (middle of top half)
Max= 89

C8. The answer is c).

C9. The answer is b).

C10. The answer is a).

C11. Standard deviation is a positive number, so the answer is a).

C12. 29, 56, 59, 62, 66, 67, 78, 81, 89

IQR= Q_3 - Q_1
First, we need the median and then the 1st and 3rd quartiles

Median=66

Q1=(59+62)/2 = 60.5

Q3=(78+81)/2 = 79.5

IQR=Q3-Q1=79.5 – 60.5 = 19

To find outliers

Q3 + 1.5 IQR= 79.5+1.5(19) = above 108, so none here

Q1 + 1.5IQR= 60.5 – 1.5(19) = below 32, so 29 is an outlier

An outlier is any number above 108 or below 32.

Therefore, there is one outlier...the number "29".

C13. 21, 23, 26, 27, 28, 32, 34, 34, 37, 38, 40, 41

$$\bar{x} = \frac{21 + 23 + \cdots + 41}{12} = 31.75$$

Make a chart to find the standard deviation.

$x - \bar{x}$	$(x - \bar{x})^2$
21-31.75	115.56
23-31.75	76.56
26-31.75	33.06
27-31.75	22.56
28-31.75	14.06
32-31.75	0.06
34-31.75	5.06
34-31.75	5.06
37-31.75	27.56
38-31.75	39.06
40-31.75	68.06
41-31.75	85.56

$$s = \sqrt{\frac{492.25}{11}} = 6.69$$

The variance is $s^2 = 44.75$

Outliers?

Q1=average of 3rd and 4th data= (26+27)/2 = 26.5

Q3=average of 9th and 10th=(37+38)/2 = 37.5

IQR= 37.5 - 26.5 = 11

Q1- 1.5(IQR)= 26.5 - 1.5(11) = 10 below

Q3+ 1.5(IQR) = 37.5 + 1.5(11) = 54 above

There are no numbers below 10 or above 54, so there are no outliers!

C14. Samuel measures in the 75th percentile means he is taller than 75% of kids his age, or shorter than only 25% of kids his age.

C15. If you replace one measurement with 60 and it was 40km, the total you are dividing by to find the mean will be 20 larger and therefore, the mean will definitely increase.

C16. IQR=Q3 - Q1
12=Q3 - 4
Q3=16

C17. If your data goes up quickly and then levels off, the tail is to the right and it would be called right skewed. The answer is b).

C18. A boxplot shows the max, min, and the quartiles, so it represents quantitative data. It would also be correct to use a histogram for graphing this data. The answer is d).

C19. If you take the mean and subtract 5lb three times, you get to 135lb. So, 150 – 3(5)=135lb and so Benjamin's weight is three standard deviations below the mean.

C20. Write the numbers in ascending order: 2, 3.5, 4.5, 5, 6, 7, 8...the median is the number in the middle, so it is 5.

C21.
n=8 data

1.5, 2.5, 4, 5.5* 6, 10, 11.5, 13
From the bottom half, the two numbers in the middle are 2.5 and 4, we average them and get (2.5+4)/2= 3.25

This is Q1=3.25

The median of the top half of the numbers is (10+11.5)/2= 10.75 which is Q3

IQR=Q3 – Q1= 10.75 – 3.25= 7.5

C22. 7,6,9,10,4,5,7,8,9,10

If you replace the 4 with a 6, the numbers would be 5,6,6,7,7,8,9,9,10,10 and the median would still be the same number and would not change. The mode would change because now, 6, 7, 9 and 10 would all be modes. The mean would be slightly larger because the total would be larger. The range would be different because the smallest number would now be 5 and not 4.

The answer is a).

C23.

$$\frac{6+2+3+5+9+x}{6} = 5$$

x=30-6-2-3-5-9=5

The missing number is 5.

C24. The IQR is not affected by outliers as it is a resistant measure. The answer is (b).

C25. It has a long tail to the left. i.e. distance from median to minimum is much greater than distance from median to max. Therefore, it is skewed to the left or neg. skewed.
The answer is (b).

C26.If the largest value is doubled, the mean would increase and the range would increase. The median or middle number would not change. But, the IQR wouldn't change either because it doesn't involve the highest value. So, c) is false.

C27. min=35 Q1=68, Median=77, Q3=83 and Max=97

The number of scores between 77 and 83 is the number of scores from the median to Q3 which is 25% of the scores, so 0.25x196=49 scores. The answer is (c).

C28. 2, 12, y,y,y,15, 18, 18, 19

Mean is 13.6666

There are 9 numbers

We can find the mean...

$$\frac{2+12+y+y+y+15+18+18+19}{9}=13.6666$$

Cross-multiply and solving for y...we get:

14+3y+70=123
3y=39
y=13
Now, the numbers are 2,12, 13, 13, 13, 15, 18, 18, 19

The median is 13. I is false

The mode is 13. II is true

Q1=12.5
Q3=18
IQR=18-12.5=6.5

An outlier would be below Q1 -1.5(IQR)= below 12.5 -1.5(6.5)=2.75...so "2" is an outlier
III is true

The answer is (b).

C29. The first quartile occurs at 0.25 xn = 0.25x50=12.5...average of 12th and 13th numbers. This is an estimate to figure out which bar you are in. If you actually write out the numbers, the Q1 would have 12 numbers below it and 12 above it and it would be the 13th number, but this is a good estimate. The first quartile occurs between 0 and 10. The answer is (a).

C30. 2A) Boxplot 3 as it is fairly uniform (rectangle)

2B) is Boxplot 2 as it is slightly right-skewed

2C) is Boxplot 4 as it is right-skewed with outliers

2D) is Boxplot 1 as it is left-skewed.

C31.
A). Graph B has a smaller IQR since the box is much thinner and in A it is much wider
The answer is (b)

B). The data is less spread out if the standard deviation is smaller, so Type B.

The answer is (b).

C). The answer is (b). 250 is too large since the min to max isn't even 250 and 3 is too small.

D). For Type B, the median is approximately 375-380. The answer is (b).

E). For Type A, more than 75% of the observations are larger than 400 since from Q1 to the max are all above 400. The answer is (d).

C32.We have 1.5+3.5+1.5+0.5=7 so an odd number of numbers so the middle number is the 4th one, so it would be in the 2nd bar since the first bar is only 1.5 numbers. So, the 25th percentile or Q1 would be between 10 and 20.

C33. Add up all of the frequencies= 3+1+4+6+4+8+2=28

First quartile= 0.25x 28=7th data...the third bar over since 3+1=4 isn't enough for the first bar

So, the first quartile is 60. Again, since there are a lot of numbers this gives us an estimate and as long as we aren't on the very last number of a bar or very first number it is good enough. Technically, you could write out the numbers 1 to 28 and know that the first 14 are in the bottom half and from 15 to 28 are in the top half, so the first quartile would be between the 7th and 8th, ie. the average of the 7th and 8th, but this is still in 3rd bar from the left, 60.

D. Normal Distribution

Density Curves

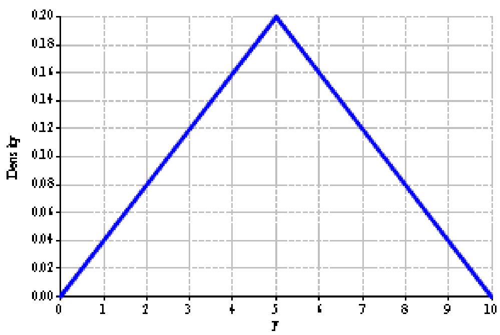
Which of the following are probability density functions?
A) no, the area under the graph isn't 1
B) yes, it is non-negative and the area is 1
C) yes, it is non-negative and the area is 1
D) no, the graph is negative from -1 to 2

Example 1.

Find the probability the salad weight is between 8 and 15 ounces.

Area=lxw=(7)(1/10)=0.7

Example 2. Given the graph, find the Pr(X<5) and Pr(X<7).

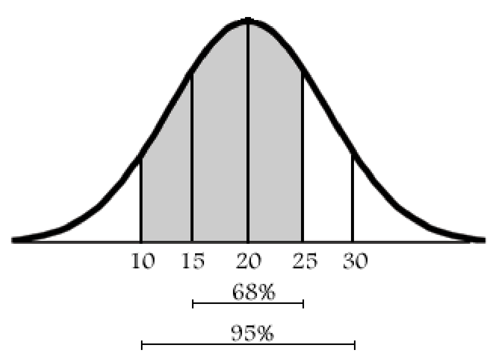


Pr(X<5)=0.50
Pr(X<7)=1- Pr(X>7)= 1- bh/2= 1- (3)(0.12)/2=0.82

Empirical Rule

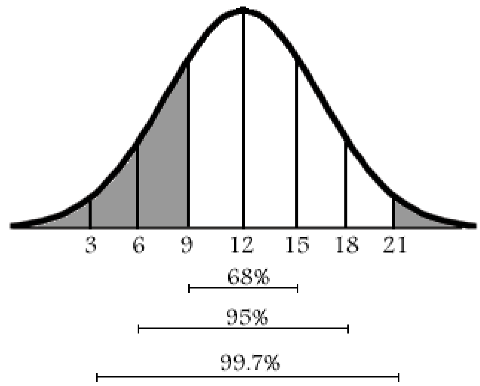
Example 1.

Mean =20 and standard dev=5
Find % between 10 and 25
See diagram to the right= 95/2 + 68/2 = 81.5%



95% lie between 10 and 30 (2 standard deviations)

Example 2.



Mean =12 and standard dev=3

Find % above 21 or below 9

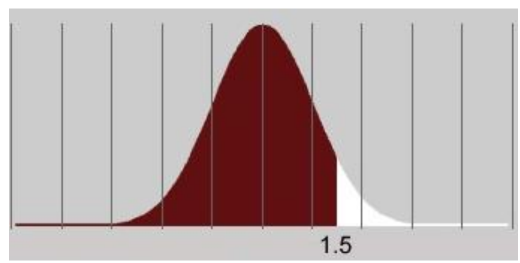
On the right, to find above 21, we see that the unshaded portion is 99.7/2%=49.85%, so the shading above 21would be 50% - 49.85%= 0.15%
On the left of the mean from 9 to 12 would be 68/2%= 34%, so the shaded area we want would be 50% - 34% = 16%
The total shading is then 0.15% + 16%=16.15%

99.7% of the data lie between 3 and 21, ie. 3 standard deviations

The Standard Normal Random Variable

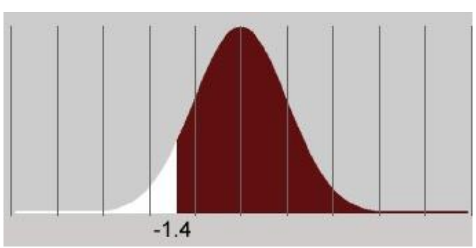
Example 1. Find each of the following probabilities by using the table for Z.

a) Pr[Z<1.5]

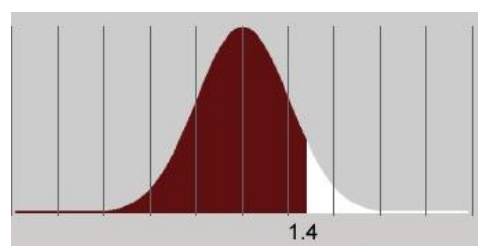


Therefore, the answer is Pr[Z<1.5]= 0.9332

b) Pr[Z>-1.4]



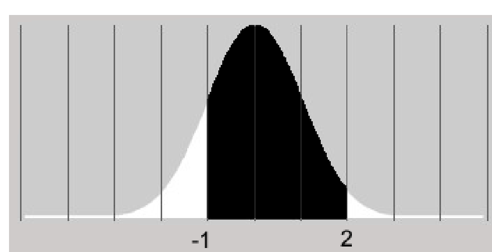
same area as



Therefore, the answer is Pr[Z>-1.4] = Pr[Z< 1.4] = 0.9192

c) Pr[-1<Z<2]

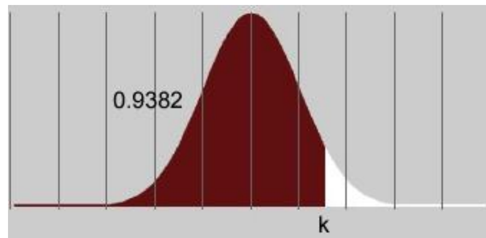
Pr[Z<2] - Pr[Z<-1] = 0.9772 - 0.1587 = 0.8185



Example 2.

Draw out 0.90 and find the value of Z, by looking up Area=0.9 in the body of the z=table. Z=1.28...the answer is (d).

Example 3. Find k such that Pr(Z<k)=0.9382



Look up the area 0.9382 and find "k" along the left side of the table.

Therefore, k=1.54

Example 4. Pr(Z>k)=0.70, so the area below Z would be 0.30

Look up the area 0.30 in the body of the Z table and you get k= -0.525.

Example 5.

a) $Z1 = \frac{x-\mu}{\sigma} = \frac{60-67}{10} = -0.70$

b) $Z2 = \frac{x-\mu}{\sigma} = \frac{80-75}{5} = 1$

c) $Z3 = \frac{x-\mu}{\sigma} = \frac{75-80}{3} = -1.67$

Z3, Z2, Z1 is largest to smallest of relative standings (largest standard deviation away from mean to smallest)

D1. For the standard normal random variable Z, find the value of Pr[Z<1.6].

A. 0.0548	B. 0.9452	C. 0.8554	D. 0.1446	E. None of the above

Pr (Z<1.6)=0.9452
The answer is b).

D2. For the standard normal random variable Z, find the value of Pr[Z>-0.80].

A. 0.7881	B. 0.80	C. 0.2119	D. 0.5319	E. None of the above

Pr(Z> - 0.80) = 1-Pr(Z<-0.80) = 0.7881
The answer is a).

D3. For the standard normal random variable Z, find the value of
Pr[-1.20<Z<1.20].

A. 2.4	B. 0.9918	C. 0.1151	D. 0.7698	E. None of the above

Pr (-1.20<Z<1.20)= Pr(Z<1.2) - Pr(Z<-1.2)
=0.8849 – 0.1151
=0.7698
The answer is d).

D4. For the standard normal random variable Z, what is the value of Pr[Z>1.76]?

A.0.0392	B. 0.9608	C. 0.9554	D. 0.0446	E. None of the above

Pr (Z> 1.76)= 1 - Pr(Z< 1.76)
= 1 - 0.9608
=0.0392
The answer is a).

D5. If Z is the standard normal random variable, find Pr[-1 < Z < 1].

A. 0.0228	B. 0.9772	C. 0.1587	D. 0.8413	E. None of the above

Pr (-1<Z<1)= Pr(Z<1) - Pr(Z<-1)
=0.8413 – 0.1587
=0.6826
The answer is e).

D6. Find the value of k if it is known that Pr[k<Z<1.5]= 0.0483 , where Z is the standard normal random variable.

A. 1.2	B. -1.2	C. 0.8849	D. 1.66	E. none of the above

Pr(Z<1.5)= 0.9332
0.9332 – 0.0483=0.8849
Look up area 0.8849 and you get k=1.2

The answer is a).

D7. Use the table for the standard normal random variable Z to find
Pr[-0.65<Z<1.92].

A. 0.6226	B. 0.2284	C. 0.7148	D. 0.2852	E. None of the above

Pr (-0.65<Z<1.92)=Pr(Z<1.92) - Pr(Z<-0.65)
=0.9726 – 0.2578
=0.7148

The answer is c).

D8. Use the table for the standard normal random variable Z to find a value of k for which Pr[Z<k]= 0.9495

A. 0.9495	B. 0.8264	C. -1.64	D. 1.64	E. None of the above

Pr(Z<k)=0.9495 same area as Pr(Z>k)

Find the area=0.9495 by looking in the body of the chart...we get k=1.64
The answer is d).

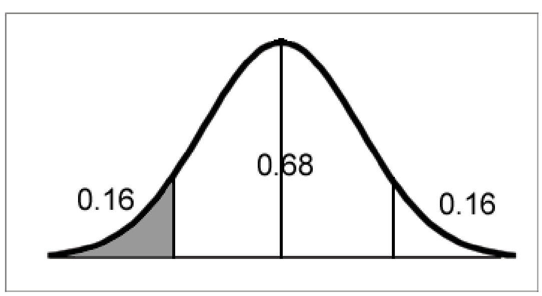
D9. Normally distributed $\mu = 75 \; and \; \sigma = 10$

Look up 0.16 area in the body and get Z_1=-0.99 and look up 0.68+0.16=0.84 area below the second Z and get Z_2=1...substitute both into the formula to find X values

$Z=\frac{X-\mu}{\sigma}$

$-0.99 = \frac{X_1-75}{10}$

$X_1 = 65.1$

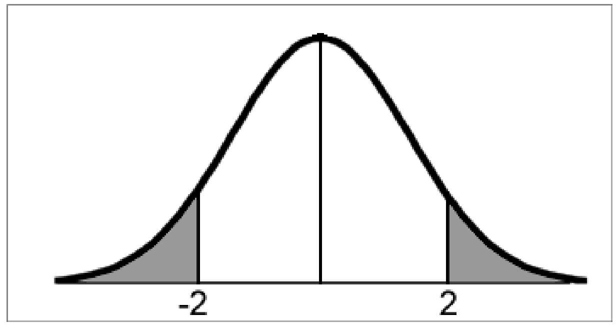


$1 = \frac{X_2 - 75}{10}$

$X_2 = 85$

The lowest mark you can get to get a C is 65.1 and if they were to ask, the highest mark would be an 85.

D10. Normally distributed $\mu = 120\ and\ \sigma = 12$



$Z = \frac{X-\mu}{\sigma}$

$$Z_1 = \frac{96-120}{12} = -2$$

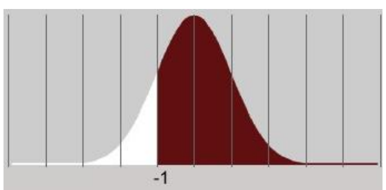
$$Z_2 = \frac{144-120}{12} = 2$$

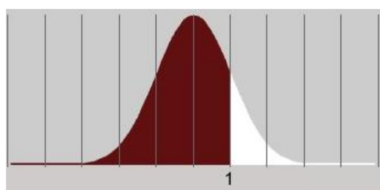
The area below 96mmHg would be the Pr(Z<-2)=0.0228 and the area above 144mmg would be Pr(Z>2)=1-Pr(Z<2)=1-0.9772=0.0228

So, the total percentage would be 0.0228x2=0.0456 or 4.56%

Normal Random Variables

Example 1. Find Pr[X>90].

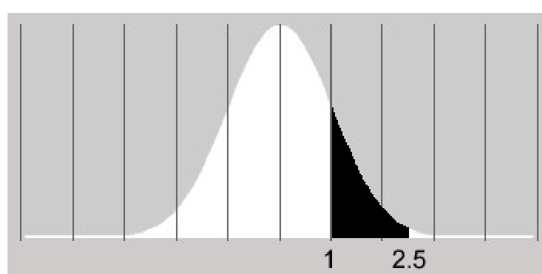




$$Z = \frac{90-100}{10} = -1$$

Pr(X>90)=Pr(Z>-1)= 1 – Pr(Z<-1)=1 – 0.1587=0.8413

Example 2. Pr[12<X<15].



Pr(12<X<15)=Pr $\left(\frac{12-10}{2} < Z < \frac{15-10}{2}\right)$
=Pr(1<Z<2.5)
=Pr(Z<2.5) - Pr(Z<1)
=0.9938 - 0.8413
=0.1525

Example 3.

Normally distributed $\mu = 200\ and\ \sigma = 15$

Pr(X>205)=Pr $\left(Z > \frac{205-200}{15}\right) = 0.33$

Pr(Z>0.33)=1-Pr(Z<0.33) = 1-0.6293=0.3707
The answer is (e).

Example 4.

Normally distributed $\mu = 200\ and\ \sigma = 15$

Look up Area=0.67 in the body of the Z-table to find the corresponding value of Z
Z=0.44

Z=$\frac{X-\mu}{\sigma}$

$$0.44 = \frac{X - 200}{15}$$

X=206.6

The answer is (b).

Example 5.

Normally distributed $\mu = 260\ and\ \sigma = 10$

Draw the Z-curve with 5% or 0.05 to the far right, and then 1-0.05=0.95 is the area to the left...

We get Z=1.645

Z=$\frac{X-\mu}{\sigma}$

$$1.645 = \frac{X - 260}{10}$$

X=276.5 days
The answer is (c).

D11.

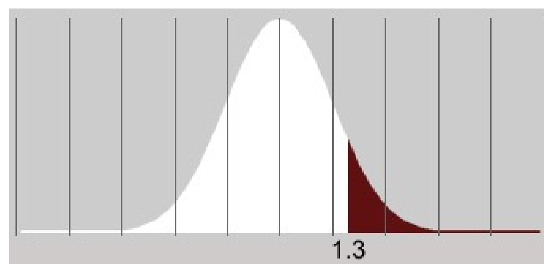
$\Pr(X<520)=\Pr(Z<\frac{520-500}{20})=\Pr(Z<1)=0.8413$

D12. X is a normal random variable with mean 35 and standard deviation 5.

$\Pr(30<X<40)=\Pr(\frac{30-35}{5}<Z<\frac{40-35}{5})=\Pr(-1<Z<1)$

=Pr(Z<1) - Pr(Z<-1)
=0.8413 - 0.1587
=0.6826

D13.



Let X be the test score. Then $X \sim N(\mu,\sigma)$ with $\mu=500$, $\sigma=100$. So

$$\Pr(X>630) = \Pr\left(Z=\frac{X-\mu}{\sigma}>\frac{630-500}{100}\right)$$
$$= \Pr(Z>1.3) = 1-\Pr(Z<1.3) = 1-0.9032 = 0.0968.$$

D14. (a) What percentage of pregnancies last less than 240 days?

Let X be the length of the pregnancy in days. Then $X \sim N(\mu,\sigma)$ with $\mu=266$, $\sigma=16$.

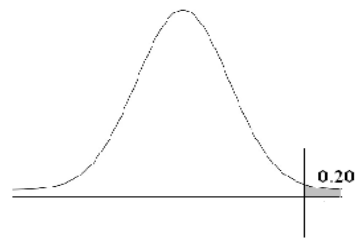
So $$\Pr(X<240) = \Pr\left(Z=\frac{X-\mu}{\sigma}<\frac{240-266}{16}\right) = \Pr(Z<-1.63) = 0.0516.$$

(b) What percentage of pregnancies last between 240 and 270 days?

$$\Pr(240 < X < 270) = \Pr\left(\frac{240-266}{16} < Z < \frac{270-266}{16}\right) = \Pr(-1.63 < Z < 0.25)$$

$$= \Pr(Z < 0.25) - \Pr(Z < -1.63) = 0.5987 - 0.0516 = 0.5471.$$

(c) How long do the longest 20% of pregnancies last? **Look up the area 0.80 in the body and get Z=0.84**



$$\Pr(Z > z) = 0.20 \quad \Rightarrow \quad z = 0.84.$$

$$z = \frac{x-\mu}{\sigma} \quad \Rightarrow \quad x = \mu + z\sigma = \mu + 0.84\sigma$$

$$= 266 + 0.84 \cdot (16) = 279.44.$$

Therefore, the longest 20% of pregnancies last more than 279 days.

D15. (a) What is the probability of getting a 91 or less on the exam?

Let X be the final grade. Then $X \sim \mathrm{N}(\mu, \sigma)$ with $\mu = 73$, $\sigma = 8$. Then

$$\Pr(X \le 91) = \Pr\left(Z = \frac{X-\mu}{\sigma} \le \frac{91-73}{8}\right) = \Pr(Z < 2.25) = 0.9878.$$

(b) What percentage of students scored between 65 and 89?

$$\Pr(65 < X < 89) = \Pr\left(\frac{65-73}{8} < Z < \frac{89-73}{8}\right) = \Pr(-1 < Z < 2)$$

$$= \Pr(Z < 2) - \Pr(Z < -1) = 0.9772 - 0.1587 = 0.8185.$$

c) Only 5% of the students taking the test scored higher than what grade?

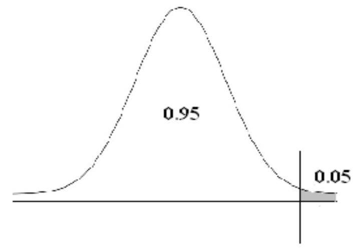
Look up the area 0.95 in the body and get Z=1.645

$\Pr(Z > z) = 0.05 \Rightarrow z = 1.645.$

$$z = \frac{x-\mu}{\sigma} \Rightarrow x = \mu + z\sigma = \mu + 1.645\sigma$$

So $x = 73 + 1.645 \cdot (8) = 86.16.$

Therefore, 5% of the students scored higher than 86%.



D16. (a) Find the probability that the monkey's weight is less than 13 pounds.

Let X be the rhesus monkey's weight in pounds. Then $X \sim N(\mu,\sigma)$ with $\mu = 15$, $\sigma = 3$.

So $$\Pr(X < 13) = \Pr\left(Z = \frac{X-\mu}{\sigma} < \frac{13-15}{3}\right) = \Pr(Z < -0.67) = 0.2514.$$

(b) Find the probability that the weight is between 13 and 17 pounds.

<u>Solution</u>:

$$\Pr(13 < X < 17) = \Pr\left(\frac{13-15}{3} < Z < \frac{17-15}{3}\right) = \Pr(-0.67 < Z < 0.67)$$

$$= \Pr(Z < 0.67) - \Pr(Z < -0.67) = 0.7486 - 0.2514 = 0.4972.$$

(c) Find the probability that the monkey's weight is more than 17 pounds.

$$\Pr(X > 17) = \Pr\left(Z > \frac{17-15}{3}\right) = \Pr(Z > 0.67) = 1 - \Pr(Z < 0.67)$$

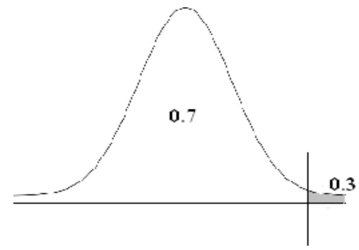
$$= 1 - 0.7486 = 0.2514.$$

D17. **(a)** What is the shortest time spent waiting for a heart transplant that would still place a patient in the top 30% of waiting times?

Let X be the waiting time (in days). Then

$X \sim N(\mu, \sigma)$ with $\mu = 127$, $\sigma = 23.5$.

So $\Pr(Z > z) = 0.30 \Rightarrow z = 0.52$.**Look up the area 0.70 in the body and get Z=0.52**



$$z = \frac{x-\mu}{\sigma} \Rightarrow x = \mu + z\sigma = \mu + 0.52\sigma$$

$$= 127 + 0.52 \cdot (23.5) = 139.2.$$

Therefore, 30% of the patients must wait for more than 139 days for a heart transplant.

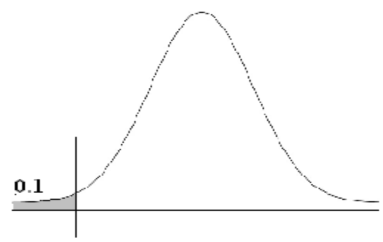
(b) What is the longest time spent waiting for a heart transplant that would still place a patient in the bottom 10% of waiting times?

Look up the area 0.10 in the body and get Z= -1.28

$\Pr(Z < z) = 0.10 \Rightarrow z = -1.28.$

$$z = \frac{x-\mu}{\sigma} \Rightarrow x = \mu + z\sigma = \mu - 1.28\sigma$$

$$= 127 - 1.28 \cdot (23.5) = 96.9.$$



Therefore, 10% of the patients have to wait for less than 97 days for a heart transplant.

D18. a) Pr(X>90)= $\Pr\left(Z > \frac{90-80}{5}\right) = \Pr(Z > 2) = 1 - \Pr(Z < 2) = 1 - 0.9772 = 0.0228$

b) Pr(X<75)=Pr $(Z < \frac{75-80}{5})$=Pr(Z<-1)=0.1587

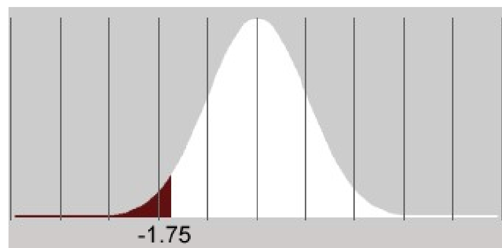
c) Pr(Z<k)=0.8665

Look up the area 0.8665 in the table and k=1.11, which means Z=1.11.

$$1.11 = \frac{X - 80}{5}$$

5.55 = X - 80
X=85.55
Therefore, the student scored approximately
85.61

D19.

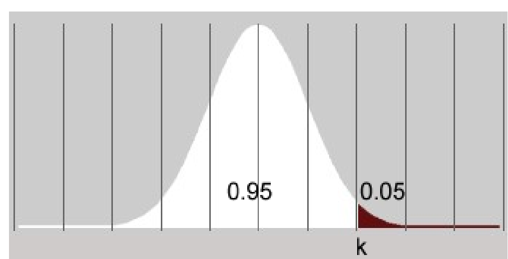


$\mu = 3700$
$\sigma = 400$

Pr(X<3000)=Pr(Z<$\frac{3000-3700}{400}$) $= \Pr(Z < -1.75) = 0.0401$

D20.

$\mu = 65$
$\sigma = 10$



Pr(Z<k)=0.95
Look up the area 0.95 on the table and k=1.645

$$Z = \frac{X - \mu}{\sigma}$$

$$1.645 = \frac{X - 65}{10}$$

X=81.5

Therefore, a student must score 81.5

D21. The answer is d).

D22. $\mu = 100$
$\sigma = 15$

Pr(X>130)=Pr(Z>$\frac{130-100}{15}$) $= \Pr(Z > 2) = 1 - \Pr(Z < 2) = 1 - 0.9772 = 0.0228$

Pr(110<X<120)=Pr($\frac{110-100}{15} < Z < \frac{120-100}{15}$) $= \Pr\ (0.67 < Z < 1.33)$
=Pr(Z<1.33) - Pr(Z<0.67)
=0.9082 - 0.7486
=0.1596

D23.
$\mu = 100$
$\sigma = 15$

$$Z = \frac{125 - 100}{15} = 1.67$$

Pr(Z<1.67)=0.9525
Therefore, she scores higher than 95% of all adults.

D24. Her height is 71 inches.
$\mu = 64$
$\sigma = 2.4$

$$Z = \frac{X - \mu}{\sigma} = \frac{71 - 64}{2.4} = 2.92$$

The Z-score is 2.92

D25.
X is a normal random variable with unknown mean μ and standard deviation σ=3. If Pr[X<25]=0.9772, what is the value of μ?

Look up the area 0.9772 in the body and you get Z=2.

$$Z = \frac{X - \mu}{\sigma}$$

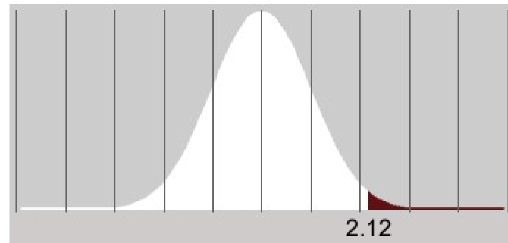
$$2 = \frac{25 - \mu}{3}$$

μ =19

D26. (a) What is the probability that a child's IQ is greater than 125?

Let X be the child's IQ. Then $X \sim N(\mu,\sigma)$ with $\mu = 100.4$, $\sigma = 11.6$. So

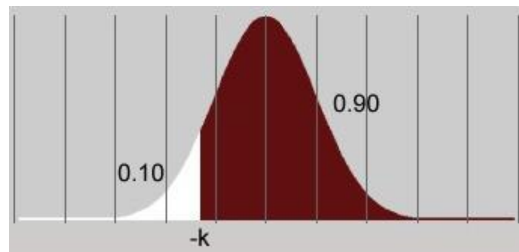
$$\Pr(X > 125) = \Pr\left(Z = \frac{X-\mu}{\sigma} > \frac{125-100.4}{11.6}\right)$$



$$= \Pr(Z > 2.12) = 1 - \Pr(Z < 2.12) = 1 - 0.9830 = 0.0170$$

(b) About 90% of the children have IQ's greater than what value?

Solution: **Look up the area 0.10 in the body and get Z= -1.28**

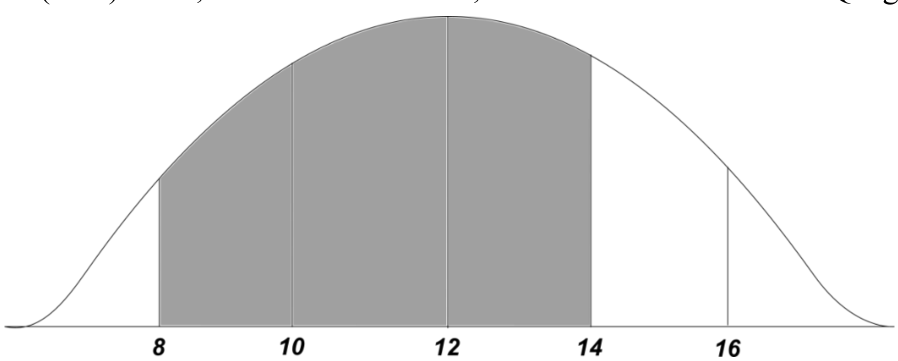


$$\Pr(Z > z) = 0.90 \Rightarrow z = -1.28.$$

$$z = \frac{x-\mu}{\sigma} \Rightarrow x = \mu + z\sigma = \mu - 1.28\sigma$$

So $x = 100.4 - 1.28 \cdot (11.6) = 85.6$.

D27. On the left from 8 to 12 is 1/2(95%)=47.5% and on the right from 12 to 14 there is 1/2(68%)=34%, so the total is 81.5%, 90% of the children have IQ's greater than 85.6.



E. Scatterplots

Example 1. (a). is true.

Example 2. (c) is true. It depends on where the outlier lies.

Example 3. A pie chart is only for studying one CATEGORICAL variable. So, (b) is false.

Example 4. (a) is appropriate, since we need two quantitative variables in order to study regression.

Example 5. (a) is the answer. It can't be above 100%, so d) is not possible.

E1. error, unitless.

E2. = -0.5... means r^2=0.25 which means 25% of the variation in the y-values can be explained by this model. The answer is (a).

E3. The answer is (ii). There is a moderately strong positive correlation.

E4. The answer is (i). The correlation between x and y is the same as the correlation between y and x.

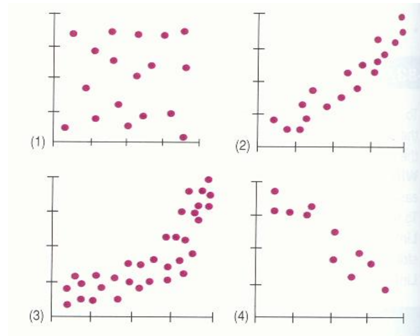
E5. The answer is (b).

E6. (a) mileage

(b) weight

(c) Since $r^2 = 0.44$ and $r < 0$, therefore $r = -\sqrt{0.44} = -0.663$. It is negative since it is a negative correlation, ie. as cars increase in weight, the miles per gallon would decrease

E7.



(a) #1 shows little or no association.

(b) #4 shows negative association. Increases in one variable are generally related to decreases in the other variable.

(c) #2 and #4 each show a linear association.

(d) #2, #3 and #4 show a moderately strong association.

(e) None shows a very strong association.

E8. (a) -0.98 (b) 0.74 (c) 0.96 (d) -0.03

E9. Correlation is not a resistant measure and it can be negative, if the slope of the line is negative. It also has the same sign as the slope. It has no units.

Therefore, the answer is b).

E10.
(b) is possible since it involves two quantitative variables and the correlation is between -1 and 1.

E11. The answer is (b).

E12. (a). is the answer since a positive relationship means as one variable increases, so does the other. Also, as one decreases, so does the other.

E13. (a) is not possible from regression line.

E14. The line goes up and to the right, so it is positive. The answer is (c). Remember, d) is impossible since correlation is between -1 and 1.

E15. The answer is (a) since r is so close to 0.

F. Regression

p.91 Use the points (0,5) and (4,2) on the line

$$slope = \frac{5-2}{0-4} = -\frac{3}{4}$$

p. 92

Q3. slope= how much the average house value increases by each year ($5632)
y-intercept= average cost of a new home in the year 1970 ($14760)

Example 1.

(a) The explanatory variable (x) is the student's ACT score, while the response variable (y) is the student's SAT score.

$$b = r\frac{s_y}{s_x} = (0.817)\frac{180}{5} = 29.412 ,$$

$$a = \bar{y} - b_1\bar{x} = 912 - (29.412)\cdot(21) = 294.348 .$$

The regression equation is $\hat{y} = a + bx = 294.348 + 29.412x$.
The answer is (ii).

(b)
$r^2 = (0.817)^2 = 0.667 = 66.7\%$.
The answer is (i).

Example 2.

x= height and $\hat{y} = foot\ length$
A). $\hat{y} = 10.9 + 0.23x$
$\hat{y} = 10.9 + 0.23(73) = 27.7\ cm$

The answer is (b).

B) Residual=observed - predicted = 29cm - 27.7 cm=1.3 cm
The answer is (b).

C) $\hat{y} = 10.9 + 0.23(70) = 27\ cm$...the answer is (a).

D) 25=10.9+0.23x solve for x
14.1=0.23x
X=61.3 inches

Example 3.

Residuals are the difference between observed and predicted responses. The answer is (c).

Example 4.

$\hat{y} = -2.3 + 1.80x$

$\hat{y} = -2.3 + 1.80(4) = 4.9$...this is the predicted y-value for x=5.

The observed value is 5.

Residual=observed - predicted = 5-4.9 = 0.1...the answer is (c).

Example 5.

Slope=b=$\frac{rs_y}{s_x}$= $0.6(\frac{2}{3}) = 0.4$ The answer is (a).

Example 6.

Intercept = a=$\bar{y} - b\bar{x} = 6 - (0.4)(5) = 4$

So, the equation is
$\hat{y} = a + bx = 4 + 0.4x$

Example 7.

The value of $r^2=0.6^2=0.36$ or 36%. The answer is (a).

Example 8.

Given: $r^2=0.8$, r=0.89

$$\bar{x} = \frac{750}{100} = 7.5$$

$$\bar{y} = \frac{525}{100} = 5.25$$

Find a and b

$$b = r\left(\frac{s_y}{s_x}\right) = 0.89\left(\frac{14.4}{12.5}\right) = 1.025$$

$$a = \bar{y} - b\bar{x} = 5.25 - 1.025(7.5) = -2.4375$$

The regression equation is $\hat{y} = a + bx = -2.4375 + 1.025x$.

Example 9.

1. Find the mean and standard deviation of x and y

$\bar{x}$ =224.1/12=18.7

$\bar{y}$=4829/12=402.4

$$s_x = \sqrt{\frac{176.99}{11}} = 4.01$$

$$s_y = \sqrt{\frac{172908.8}{11}} = 125.4$$

I	x_i	y_i	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$\frac{x_i-\bar{x}}{s_x}$	$\frac{y_i-\bar{y}}{s_y}$	$\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$
1	14.2	215	20.25	35118.76	-1.12	-1.49	1.67
2	16.4	325	5.29	5990.76	-0.57	-0.61	0.35
3	11.9	185	46.24	47262.76	-1.7	-1.72	2.92
4	15.2	332	12.25	4956.16	-0.87	-0.56	0.49
5	18.5	406	0.04	12.96	-0.05	0.03	0
6	22.1	522	11.56	14304.16	0.85	0.95	0.81
7	19.4	412	0.49	92.16	0.17	0.08	0.01
8	25.1	614	40.96	44774.56	1.6	1.68	2.69
9	23.4	544	22.09	20050.56	1.17	1.12	1.31
10	18.1	421	0.36	345.96	-0.15	0.15	-0.02
11	22.6	445	15.21	1814.76	0.97	0.34	0.33
12	17.2	408	2.25	31.36	-0.37	0.04	-0.01
Total	224.1	4829	176.99	172908.8			10.55

2. Calculate the correlation coefficient using the formula

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

$$r = \frac{1}{11}(10.55) = 0.96$$

3. Find the intercept and the slope and explain what they mean.

$$b = r\frac{s_y}{s_x} = 0.96\frac{125.4}{4.01} = 30.02$$

$$a = \bar{y} - b\bar{x} = 402.4 - 30.02(18.7) = -159$$

The y-intercept is the Ice Cream sales when the temperature is 0 and the slope is the increase in sales in dollar for every one-degree Celsius increase in temperature. Here, the sales increases by $30 for every increase in temperature by one-degree Celsius.

4. Write the equation of the least-squares regression line

$$\hat{y} = a + bx = -159 + 30.02x$$

5. Use #4 to predict the Temperature when the Sales are $350.

350= -159+30.02x
30.02x=509
x=16.96

6. The fraction of the variation in the y-values that is explained by the regression is________.
0.96^2=0.92

7. $\hat{y} = a + bx = -159 + 30.02x$ substitute x=39
$\hat{y} = -159 + 30.02(39) = \$1011.78$

This involves extrapolation as our data only goes up to about 25 ^{0}C, so it is unlikely it will be accurate.

<u>Example 10</u>

$\hat{y} = 3.5 - 0.62x$
$s_x = 3 \ and \ s_y = 3.6$

$$b = r\frac{s_y}{s_x}$$

$$-0.62 = r\left(\frac{3.6}{3}\right)$$

r= - 0.52

Example 11.

$$r=\frac{(1.35)(-0.96)+0.6(1.11)+(0.2)(-0.67)+(-1.04)(-0.57)+(-1.11)(1.09)}{4} = -\frac{1.3811}{4} = -0.345$$

Example 12. Scatterplot

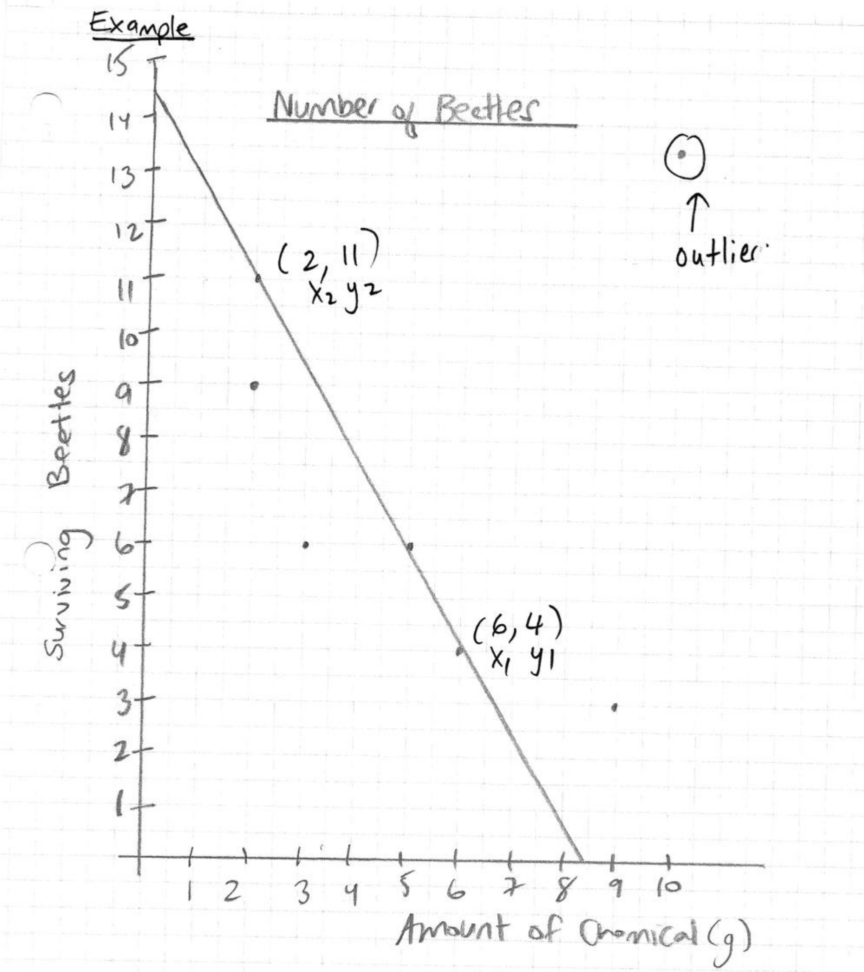
a) The explanatory variable is the amount of chemical being applied in grams. The response variable is the number of surviving beetles.

b) $b = \frac{y2-y1}{x2-x1} = \frac{11-4}{2-6} = -\frac{7}{4}$ (slope)

y-intercept is 14.5 (where the graph crosses the x axis)

Equation would be $\hat{y} = 14.5 - \frac{7}{4}x$

c) There is an outlier at (10,14).



d) 4.2 g (answers vary based on your graph)

e) $\hat{y} = 14.5 - \frac{7}{4}x$ let x=4 and solve for y

$$\hat{y} = 14.5 - \frac{7}{4}x(4)$$
$$\hat{y} = 7$$

So, there would be 7 surviving beetles.

F1. $b = r\frac{s_y}{s_x} = r\left(\frac{1}{1}\right) = r$, so, the correlation will be the same as the slope of the least-squares regression line. The answer is d).

F2. (a) The answer is (v). The slope represents how much the response variable (Winning %) changes due to an increase in the explanatory variable (Goals Allowed) of one unit.

(b) If $x = 251$, then $\hat{y} = 116.95 - 0.26 \cdot (251) = 51.69$.

(c) The answer is (ii).

F3. (a) True. It is a negative relationship, so more expensive cars will have lower fuel efficiency.

(b) True...r=-0.3 means there is a linear relationship, so it is a moderately straight line.
The correlation coefficient isn't close to -1, but much closer to 0, so it is a fairly week relationship.

(c) False. Correlation doesn't tell us about outliers.

(d) False. Correlation has no units and it doesn’t change when units are changed.

F4. (a) If the car you are thinking of buying has a 200- horsepower engine, what does this model suggest your gas mileage would be?.

$$mpg = 46.87 - 0.084HP = 46.87 - 0.084(200) = 30.07 \text{ mpg}.$$

(b) Explain what the slope means in the context.

According to the model, slope $= -0.084$ means that as horsepower increases by 1 HP, we expect mpg to go down by 0.084.

F5. Fill in the missing information in the table below. Show your work.

	$\bar{x}$	s_x	$\bar{y}$	s_y	r	$\hat{y} = a + bx$
(a)	30	4	18	6	-0.2	
(b)	100	18	60	10	0.9	

Solution:

	$\bar{x}$	s_x	$\bar{y}$	s_y	r	$\hat{y} = a + bx$
(a)	30	4	18	6	-0.2	$\hat{y} = 27 - 0.3x$
(b)	100	18	60	10	0.9	$\hat{y} = 10 + 0.5x$

(a) $b = r\frac{s_y}{s_x} = (-0.2)\frac{6}{4} = -0.3$, $a = \bar{y} - b\bar{x} = 18 - (-0.3)\cdot(30) = 27$,

$\hat{y} = a + bx = 27 - 0.3x$

(b) $b = r\frac{s_y}{s_x} = (0.9)\frac{10}{18} = 0.5$, $a = \bar{y} - b\bar{x} = 60 - (0.5)\cdot(100) = 10$,

$\hat{y} = a + bx = 10 + 0.5x$

F6. Dependent variable is: Home Attendance
R-squared = 48.5%
Variable Coefficient
Constant -14364.5
Wins 538.915

(a) Write the equation of the regression line.
Solution:
$\widehat{\text{Attendance}} = -14364.5 + 538.915(\text{Wins})$

(b) Estimate the Average Attendance for a team with 50 Wins.
Solution:
$\widehat{\text{Attendance}} = -143645 + 538915(50) =$ 12581. (Note: This is an extrapolation.)

(c) Interpret the meaning of the slope of the regression line in this context.
Solution:
For each additional win, the model predicts an increase in attendance of 538.915 people on average.

(d) In general, what would a negative residual mean in this context?
Solution:
A negative residual means that the team's actual attendance is lower than the attendance model predicts for a team with as many wins.

(e) The St. Louis Cardinals, the 2006 World Champions, are not included in these data because they are a National League team. During the 2006 regular season, the Cardinals won 83 games and averaged 42,588 fans at their home games. Calculate the residual for this team, and explain what it means.

Solution:

$$\widehat{\text{Attendance}} = -14364.5 + 538.915(83) = 30{,}365.445.$$

$$\text{Residual} = \text{observed - predicted} = 42{,}588 - 30{,}365.445 = 12{,}222.555.$$

The large positive residual shows that home attendance for the St. Louis Cardinals was much higher than is predicted according to the regression line for American League attendance

F7. Answers will vary.

Some examples include: whether or not mothers took vitamins and ate healthy during pregnancy, nutrition of child during first four years, whether or not illegal drugs were consumed, stimulation during the first four years, smoking during pregnancy, etc.

F8. a) Find the slope and intercept of the regression line.

Let X=women and Y= men

Then, we are given:

$\bar{x} = 64$

$\bar{y} = 69.3$

$s_x = 2.7$

$s_y = 2.8$

r=0.6

Slope= b= $r\frac{s_y}{s_x} = (0.6)\frac{2.8}{2.7} = 0.62$

Intercept a=$\bar{y} - b\bar{x} = 69.3 - 0.62(64) = 29.62$

b) Find the equation of the least-squares regression.

$\hat{y} = a + bx = 29.62 + 0.62x$

F9.

$\bar{x} = \frac{900}{50} = 18$

$\bar{y} = \frac{750}{50} = 15$

$s_x = 10.5$

$s_y = 12.5$

r=$\sqrt{0.92} = 0.959$

Slope= b= $r\frac{s_y}{s_x} = (0.959)\frac{12.5}{10.5} = 1.14$

Intercept a=$\bar{y} - b\bar{x} = 15 - 1.14(18) = -5.52$

c) Find the equation of the least-squares regression.

$\hat{y} = a + bx = -5.52 + 1.14x$

F10.

I	x_i	y_i	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$\frac{x_i - \bar{x}}{s_x}$	$\frac{y_i - \bar{y}}{s_y}$	$\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$
1	5	10	1	11.56	-0.63	-0.88	0.55
2	4	9	4	19.36	-1.27	-1.14	1.45
3	7	16	1	6.76	0.63	0.68	0.43
4	6	14	0	0.36	0	0.16	0
5	8	18	4	21.16	1.27	1.19	1.51
Total	30	67	10	59.2			3.94

Find the mean and standard deviation of x and y

$\bar{x} = 30/5=6$

$\bar{y}=67/5=13.4$

$s_x = \sqrt{10/4}=\sqrt{2.5} = 1.58$

$s_y = \sqrt{59.2/4} = \sqrt{14.8} = 3.85$

2. Complete the chart above and then calculate the correlation coefficient using the formula $r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$ ...graph

$r = \frac{3.94}{4} = 0.99$

3. Find the intercept and the slope

Slope= b= $r\frac{s_y}{s_x} = \frac{(0.99)3.85}{1.58} = 2.41$

Intercept a=$\bar{y} - b\bar{x} = 13.4 - (2.41)(6) = -1.06$

4. Write the equation of the least-squares regression line

$\hat{y} = a + bx = -1.06 + 2.41x$

5. Use #4 to predict the y-value when x=5.5.

$\hat{y} = a + bx = -1.06 + 2.41(5.5) = 12.2$

6. The fraction of the variation in the y-values that is explained by the regression is________.
r^2=0.98

F11.
A) The slope is -0.002, the number in front of the independent or explanatory variable. The answer is (b).

B). Time= 3.80 - 0.002 Thrust

60=3.80 - 0.002 x
60-3.80 = - 0.002x

x= -28 100 HP

F12. The slope is the number in front of the "midterm" mark. The answer is (c). If you score 10 points higher on the midterm, it would be 0.5 (10)=5 points higher (positive change of 5 pts) on the final exam. A) is not an interpretation of slope, so read the question carefully!

F13.

X	500	1000	1500	2000	2500
y	50	100	150	200	250

As x increases, we can see that y increases, so it is a positive relationship. Since these points would come close to a straight line. As x goes from 500 to 1000, x goes up by 50 and as x goes from 1000 to 1500, y goes up by 50 again. So, y goes up 50 each time as x goes up by 500.! So, it is a very strong linear relationship. If you look closely, as x goes up each time by 50, y goes up by exactly 1000 each time, so it would be a perfect straight line and r would be 1. The answer is (c).

F14. The answer is (a).

F15.

I	x_i	y_i	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$\frac{x_i-\bar{x}}{s_x}$ $\frac{y_i-\bar{y}}{s_y}$	$\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$
1	45	275	538.24	3552.16	1.39 -0.67	-0.93789
2	12	401	96.04	4408.96	-0.587 0.7517	-0.4413
3	3	420	353.44	7293.16	-1.126 0.9668	-1.08865
4	17	212	23.04	15030.76	-0.2876 -1.38798	0.39918
5	32	365	104.04	924.16	0.6111 0.34416	0.2103
Total	109	1673	1114.8	31209.2		-1.85836

Find the mean and standard deviation of x and y

$$\bar{x} = \frac{109}{5} = 21.8$$

$\bar{y}$=1673/5=334.6

$s_x = \sqrt{1114.8/4}$=$\sqrt{278.7} = 16.69$
$s_y = \sqrt{31209.2/4} = \sqrt{7802.3} = 88.33$
2. Complete the chart above and then calculate the correlation coefficient using the formula
$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$ ...graph

$$r = \frac{-1.85836}{4} = -0.46$$

F16.
(a) is false as the car gets older by one year, the selling price drops by 1.2 x1000=$1200

(b) is false, since the drop is in dollars, not in percent

(c) is false, as the new car at age=0 would be 25 x1000=$25000
(d) is false, same as (c)

The answer is (e).

F17. H= -1.3 + 1.5C
Substitute C=65

H= -1.3 + 1.5(65)=96.2 ft

F18. The answer is (a).

F19. The answer is (b).

F20. The answer is (c).

F21. The answer is (e).

F22.
Find r, a and b.

Given: r=0.76

Find a and b

$$b = r\frac{s_y}{s_x} = (0.76)\frac{1.9}{2.4} = 0.60$$
$$a = \bar{y} - b\bar{x} = 40 - 0.60(60) = 4$$
a=4
Find the least-squares regression line equation.

$$\hat{y} = a + bx = 4 + 0.6x$$

F23. It says “from the height of his wife”, so the x= wife’s height
$\bar{x} = 64.5\ and\ \bar{y} = 68.5$
$s_x = 2.5\ and\ s_y = 2.7$

$r = \sqrt{0.25} = 0.5$

A)
$b = r\frac{s_y}{s_x} = 0.5\left(\frac{2.7}{2.5}\right) = 0.54$ The answer is d).

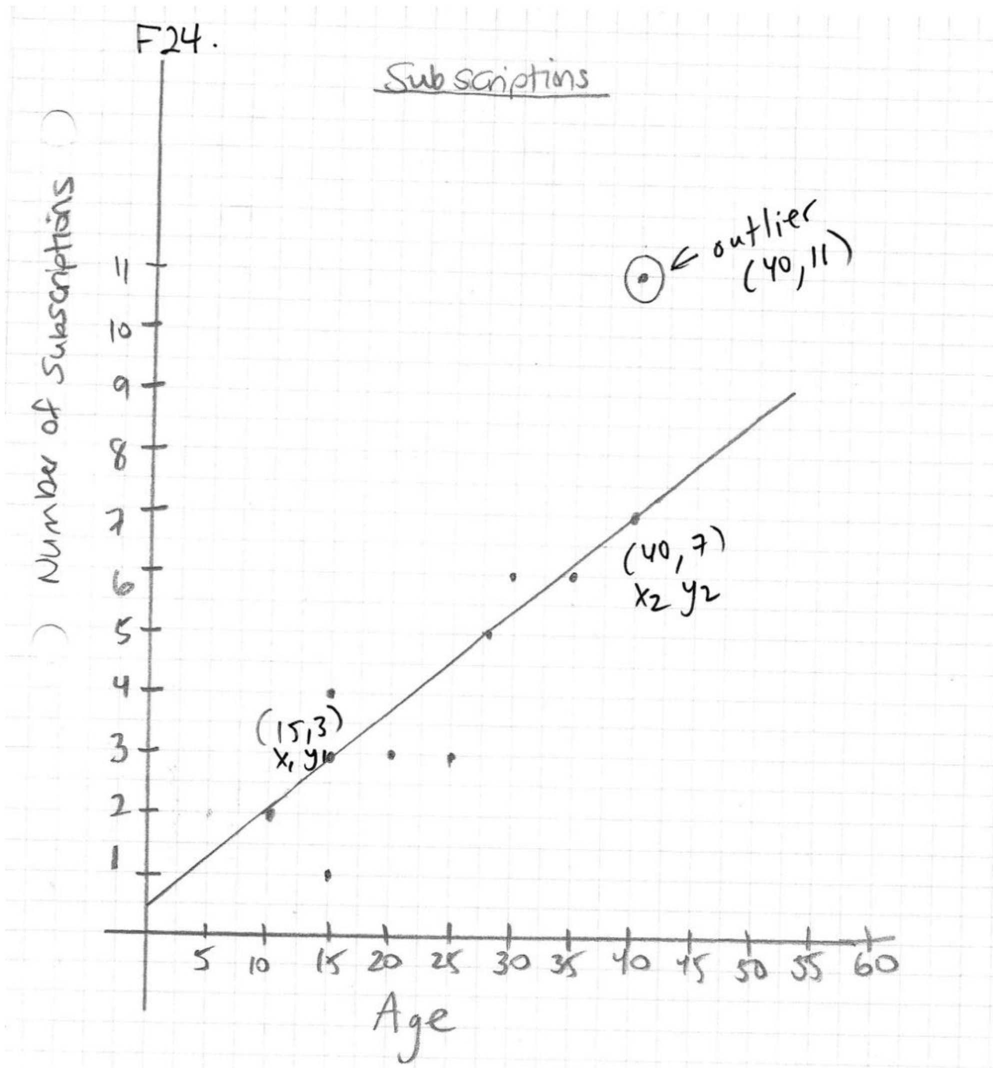
B) $\hat{y} = a + bx$
$a = \bar{y} - b\bar{x}$= 68.5 - 0.54(64.5) = 33.67

$\hat{y} = 33.67 + 0.54x$ substitute x=67
$\hat{y} = 33.67 + 0.54(67) = 69.85\ inches$

C) $b = r\frac{s_y}{s_x}$=$0.5\left(\frac{2.5}{2.7}\right) = 0.46$
F24. $b = \frac{y2-y1}{x2-x1} = \frac{7-3}{40-15} = \frac{4}{25}$ (slope)

y-intercept is 0.5 (where the graph crosses the x axis)

The equation is $\hat{y} = 0.5 + \frac{4}{25}x$



G. Two-Way Tables

Marginal Distribution

Somewhat impressed= 90/765 (100)=11.8%
Indifferent= 95/765 (100)=12.7%
Very Impressed= 350/765 (100)=45.8%

Conditional Distributions

Females somewhat impressed=50/420 (100)=11.9%
Males somewhat impressed= 40/345 (100)=11.6%
% of these who are indifferent were male? 50/95 = 0.526 or 52.6%

Example 2.

a) 13/36=0.36 or 36%

b) 13/52= 0.25 or 25%

c) 15/100=15%

Example 3.

The answer is (c).

Example 4.

Use the following table to find the percentage below:

Class	High quality	Medium Quality	Low Quality	Total
Freshman	60	20	20	100
Sophomore	50	30	40	120
Junior	60	40	70	170
Senior	30	60	70	160
Total	200	150	200	550

a) Of the students who felt campus residences are high quality, what percent are juniors?

Pr(Junior/High quality)=
$=\frac{60}{200} = 0.3\ or\ 30\%$

b) Pr(Not low quality/sophomores)=$\frac{50+30}{120} = \frac{80}{120} = 0.67\ or\ 67\%$

c) Pr(not freshman/medium) $=\frac{130}{150}=0.86666\ or\ 86.7\%$

Example 5.

Use the following table to answer the question below:

Age group	Female	Male	Total
15 to 17 years old	100	150	250
18 to 24 years old	5000	4500	9500
25 to 34 years old	1800	1500	3300
35 years or older	1500	900	2400
Total	8400	7050	15450

a) What is the probability that the selected student is 18 to 24 years old? 9500/15450=0.61 or 61%

b) What is the probability that a randomly selected female is 35 years or older?

$=\frac{1500}{8400}=0.18\ or\ 18\%$

c) What is the probability a randomly selected male is over 24 years old?
(1800 + 1500)/ 7050=0.468 or 46.8 %

d) If you randomly select a student who is 35 and over, what is the probability they are female?
1500/2400=0.625

Example 6. See the table below:

Age Groups	**Fail/Success**	**Treatment A**	**Treatment B**	**Total**
<40	Fail	5	35	40
	Success	80	235	315
$40+$	Fail	70	25	95
	Success	190	50	240

Combined data:

Fail/Success	**Treatment A**	**Treatment B**	**Total**
Fail	75	60	135
Success	270	285	555

a) Calculate the success rates for treatments A and B when the data is split by age groups. Which treatment is better?

Treatment A

<40 Success $=\frac{80}{85}=\boxed{0.941}$

$40+$ Success $=\frac{190}{190+70}=\frac{190}{260}=\boxed{0.731}$

Treatment B

< 40 Success $= \frac{235}{35+235} = \frac{235}{270} = \boxed{0.870}$

$40 +$ Success $= \frac{50}{25+50} = \frac{50}{75} = \boxed{0.667}$

∴ the success rate is higher in both age groups for treatment A than treatment B

b) Calculate the success rates for treatments A and B when the data is combined. Which treatment has a higher success rate?

Combined Treatment A Treatment B

Success rate $= \frac{270}{270+75} = \boxed{0.783}$ $= \frac{285}{60+285} = \frac{285}{345} = \boxed{0.826}$

∴ when we combine the data, treatment B has a higher success rate than treatment A

c) From a) and b) is this an example of Simpson's Paradox? Why or why not?

Yes, this is an example of Simpson's Paradox because when the data was separated by age groups, Treatment A had a higher success rate for each age group. However, once the data was combined, Treatment B has a higher success rate. When the relationship reverses when the data is combined, this is what is referred to as Simpson's Paradox.

Example 7. A population contains 1000 individuals, of which 300 carry the gene for a disease. Equivalent ways to express this proportion are as follows:

30 % of all individuals carry the gene
The proportion who carry the gene is 0.30
The probability that someone carries the gene is 0.30
The risk of carrying the gene is 0.30
The odds of carrying the gene are 300 to 700 or 3:7

Example 8.

If we have a hypothetical group of smokers (exposed) and non-smokers (not exposed), then we can look for the rate of lung cancer (event). If 20 smokers have lung cancer, 85 smokers do not have lung cancer, 3 non-smokes have lung cancer, and 99 non-smokers do not have lung cancer, the odds ratio is calculated as follows.

First, we calculate the odds in the exposed group.

- Odds in exposed group = (smokers with lung cancer) / (smokers without lung cancer) = 20/85 = 0.235

Next, we calculate the odds for the non-exposed group.

- Odds in not exposed group = (non-smokers with lung cancer) / (non-smokers without lung cancer) = 3/99 = 0.03

Finally, we can calculate the odds ratio.

- Odds ratio = (odds in exposed group) / (odds in not exposed group) = 0.235 / 0.03 = 7.8

  You can also do ad/bc=(20)(99)/(85)(3)=1980/255=7.8

Thus, using the odds ratio, this hypothetical group of smokers has approximately 8 times the odds of having lung cancer than non-smokers.

Example 8. continued

If we have a hypothetical group of smokers (exposed) and non-smokers (not exposed), then we can look for the rate of lung cancer (event). If 20 smokers have lung cancer, 85 smokers do not have lung cancer, 3 non-smokes have lung cancer, and 99 non-smokers do not have lung cancer, the relative risk ratio is calculated as follows.

Exposed	Lung Cancer	No Lung Cancer
Yes	20 a	85 b
No	3 c	99 d

Relative Risk $= \frac{a(c+d)}{c(a+b)} = \frac{20(3+99)}{3(20+85)} = \frac{2040}{315} = 6.5$

The relative risk for smokers developing lung cancer is 6.5 times that of non-smokers developing lung cancer.

Example 9.

If we hypothetically find that 18% of smokers develop lung cancer and 2% of non-smokers develop lung cancer, then we can calculate the relative risk of lung cancer in smokers versus non-smokers as:

Relative Risk = 18% /2% = 9

Thus, smokers are 9 times more likely to develop lung cancer than non-smokers.

Example 10. Calculate the Odd's Ratio.

Runs more that 25km/week		Experienced Joint Pain: No	Yes	Total
No	Count	215	75	290
	% of Non-runners	74%	26%	100%
Yes	Count	785	380	1165
	% of Runners	67%	33%	100%
Total	Count	1000	455	1455

The odds that a runner has joint pain: 380/785=0.484

The odds that a non-runner has joint pain:75/215=0.349

Odds Ratio= 0.484/0.349=1.39

a= 380 (exposed and has joint pain)

b= 785 (exposed and no joint pain)

c= 75 (not exposed and has joint pain)

d= 215 (not exposed and no joint pain)

OR use Odds Ratio= $\frac{ad}{bc} = \frac{380(215)}{785(75)} = 1.39$

Practice Exam Questions

G1. The following two-way table shows the age and sex of all undergraduate university students at a particular university.

Age Group	Female	Male	Total
15-17 years	200	250	450
18-20	3000	3500	6500
21-26	2000	2500	4500
27-34	800	900	1700
35+	500	300	800
Total	6500	7450	13950

a) How many university undergraduates are there at this university?13950
b) Find the marginal distribution of age group.

15-17 years= $\frac{450}{13950} \times 100 = 3.2\%$

18-20 years= $\frac{6500}{13950} \times 100 = 46.6\%$
etc.

c) Find the conditional distribution of females age 21-26

$$\frac{2000}{6500} \times 100 = 30.8\%$$

G2. Given the following two-way table, answer the questions below.
University students were asked how likely they think it will be that they earn a 6-digit salary in the next 20 years.

Opinion	Female	Male	Total
Almost no chance	300	50	350
Some chance, but not likely	400	300	700
A 50-50 chance	500	400	900
A good chance	400	700	1100
Almost certain	100	300	400
Total	1700	1750	3450

a) How many individuals are described using this table?

3450

b) How many males are among those surveyed?

1750

c) Find the percent of females among the respondents.

$$\frac{1700}{3450} \times 100 = 49.3\%$$

d) Does part c) represent a marginal or conditional distribution? Why?

It represents the marginal distribution of sex.

e) What percent of females thought they had a good chance to earn 6-figures in the next twenty years?

$$\frac{400}{1700} \times 100 = 23.5\%$$

f) Does part e) represent a marginal or conditional distribution? Why?
The conditional distribution of chance to earn 6-figures among females.

G3. Show that the following data is an example of Simpson's Paradox.

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

This is a real-life example from data of the University of California, Berkeley. They were sued for bias against women who had applied for admission to graduate schools there. If you look at the total data for applicants admitted, you get the table below:

	Applicants	Admitted
Men	8442	44%
Women	4321	35%

When you look at the chart above and examine individual departments, however, there is no significant bias against women. It appears sometimes the women applied in cases where very few applicants would be admitted.

This is an example of Simpson's Paradox.

G4.
a) How many employees are there in this company? 420
b) What percentage of employees are in management? $\frac{90}{420} \times 100 = 21.4\%$
c) What type of distribution does your answer to part b) represent?

The marginal distribution of employees.

d) What percentage of employees take a car?
$$\frac{82}{420} \times 100 = 19.5\%$$

e) What type of distribution does your answer to part d) represent?

The marginal distribution of mode of transportation.

f) What percentage of management take a train?
$$\frac{44}{90} \times 100 = 48.9\%$$

g) What type of distribution does your answer to f) represent?

G5. 30/100 or 0.30

G6. Conditional distribution of origin for staff?

American=90/170=52.9%
European=50/170=29.4%
Asian=30/170=17.6%

*G7. Given the table below, calculate the odds ratio and the Relative Risk:

First Child at Age 25 or Older	Breast Cancer	No Breast Cancer
YES	30	1590
NO	65	4480

Odds ratio = (odds in exposed group) / (odds in not exposed group)

$$= \frac{ad}{bc} = \frac{30(4480)}{1590(65)} = \frac{134400}{103350} = 1.3$$

Therefore, the odds of developing breast cancer is 1.3times greater for women who had their first child at 25 or older.

Relative Risk $= \frac{a(c+d)}{c(a+b)} = \frac{30(65+4480)}{65(30+1590)} = \frac{136350}{105300} = 1.29$

Therefore, the risk of developing breast cancer is 1.29 times greater for women who had their first child at 25 or older.

G8.
a) Car=6/10=60%
Train= 2/10=20%
Plane=2/10=20%
b) (2+1)/6 = 3/6 = 50%
c) 12/16 = 0.75 or 75%

G9. % of males that are liberal=35/90 = 0.388 or 39%

G10. See the table below:

Age Groups	Fail/Success	Treatment A	Treatment B	Total
< 40	Fail	10	40	50
	Success	80	235	315
$40 +$	Fail	80	30	110
	Success	190	50	240

Combined the data:

Fail/Success	Treatment A	Treatment B	Total
Fail	90	70	160
Success	270	285	555

a) Calculate the success rates for treatments A and B when the data is split by age groups. Which treatment is better?

Treatment A

< 40 Success $= \frac{80}{10+80} = \frac{80}{90} = \boxed{0.889}$

$40 +$ Success $= \frac{190}{80+190} = \frac{190}{270} = \boxed{0.704}$

Treatment B

< 40 Success $= \frac{235}{235+40} = \frac{235}{275} = \boxed{0.855}$

$40 +$ Success $= \frac{50}{30+50} = \frac{50}{80} = \boxed{0.625}$

$\therefore$ the success rate is higher in both age groups for treatment A than treatment B

b) Calculate the success rates for treatments A and B when the data is combined. Which treatment has a higher success rate?

Combined — Treatment A — Treatment B

Success rate $= \frac{270}{90+270} = \frac{270}{360} = \boxed{0.75}$ $= \frac{285}{70+285} = \frac{285}{355} = \boxed{0.803}$

$\therefore$ when we combine the data, treatment B has a higher success rate than treatment A

c) From a) and b) is this an example of Simpson's Paradox? Why or why not?

Yes, this is an example of Simpson's Paradox because when the data was separated by age groups, Treatment A had a higher success rate for each age group. However, once the data was combined, Treatment B has a higher success rate. When the relationship reverses when the data is combined, this is what is referred to as Simpson's Paradox.

H. Basic Concepts Practice Exam #1

H1. 68% are within 1 standard deviation. So, (7.5-1.4, 7.5+1.4)=(6.1,8.9) So, 68% lie between 6.1 cm and 8.9 cm

H2. See Q#H1. Find the first quartile for the length of the pine needle.

Look up the area=0.25 and you get a z-value of z=-0.675

Sub into the z-formula: $z = \frac{x-\mu}{\sigma}$ $\quad -0.675 = \frac{x-7.5}{1.4}$ x=6.56

The first quartile is 6.6cm.

H3. Pr(Z<1.6)- Pr(Z<-0.45)=0.9452-0.3264=0.6188

H4. Look up the area =0.90 because of it is in the top 10% then there is an area of 90% below it.

Z=1.28 and sub into Z-formula $z = \frac{x-\mu}{\sigma}$ $\quad 1.28 = \frac{x-1000}{15}$ x=1019.2 mL

H5. e) is false because r=0 doesn't mean there is no relationship. R only talks about the linear relationships that exist between x and y values.

H6. as x increases, y also increases, so it is a positive relationship.

H7. $Z = \frac{x-\mu}{\sigma} = \frac{95-75}{4} = \frac{20}{4} = 5$. Tina's mark is 5 standard deviations ABOVE the mean.

H8. n=15…write out data

22, 24, 27, 31, 33, 36, 38,* 40*, 51, 52, 54, 61, 64, 66, 67

Median = middle number = 40

Q1= middle of bottom half= 31
Q3= middle of top half= 61

IQR=Q3 – Q1= 61 – 31 = 30

H9. The answer is c). because if we have large data sets, we should group data and use a histogram.

H10. The answer is b). because all of them are quantitative except ID number and blood type.

H11. Given n=100 students and the data:

$$\sum x_i = 7500, \sum y_i = 525, s_x = 12.5,\ \ s_y = 14.1\ and\ r^2 = 0.86,\ \ find:$$

a) the intercept
b) the slope
c) the equation for the least-squares line

Solution:

r=$\sqrt{0.86} = 0.93$

$$slope = b = r\frac{s_y}{s_x} = 0.93\left(\frac{14.1}{12.5}\right) = 1.05$$

$$a = \overline{y} - b\overline{x} = 5.25 - 1.05(75) = -73.5$$

Equation of the least squares regression line $\hat{y} = a + bx = -73.5 + 1.05x$

H12. Put the data in increasing order first…

23, 23, 45, 45, 56, 77, 77, 85, 87, 90, 100

Median=77
Q1= middle of bottom half=45
Q3= middle of top half=87

IQR=Q3-Q1=87 – 45 = 42

H13. See the data from #H12. Are there any outliers?

Solution:

Outliers occur below Q1-1.5(IQR) and above Q3+1.5(IQR)

Q1-1.5(IQR)= 45 – 1.5(42)=-18 and there are no data below -18

Q3+1.5(IQR)= 87 + 1.5(42)= 150 and there are no data above 150

So, there are no outliers.

H14. The answer is d). The others are all used for quantitative variables

H15. 25% lie between the min. and the first quartile and 50% lie between the first and third quartiles.

H16. The standard deviation has the same units as the mean, so the answer is b).

H17. 95% of the data lies between two standard deviations of the mean...

65-2(3.5)=58

65+2(3.5)=72
Therefore, 95% of the bean plants lie between 58 cm and 72 cm.

H18. Find the standard deviation of the values: 3.2, 3.4, 3.0, 4.5, 4.2.

$$\bar{x} = \frac{3.2 + \cdots + 4.2}{5} = 3.66$$

$$s = \sqrt{\frac{(3.2 - 3.66)^2 + \cdots + (4.2 - 3.66)^2}{4}} = 0.65$$

H19. Find the interquartile range for the standard normal distribution.

Q1= look up area = 0.25 and we get z=-0.675

Q3= look up area 0.75 and we get z= 0.675

IQR=Q3-Q1=0.675 –(-0.675)=1.35

H20. Use the z-formula with z=1.7 and $\mu = 100\ and\ \sigma = 15$

$z = \frac{x-\mu}{\sigma}$ $1.7 = \frac{x-100}{15}$ x=125.5

H21. See #H20.

Find Pr(95<X<110)
$=\Pr\left(\frac{95-100}{15} < Z < \frac{110-100}{15}\right) = \Pr(-0.33 < Z < 0.67) = \Pr(Z < 0.67) - \Pr(Z < -0.33)$
=0.7486 – 0.3707
=0.3779

H22. See #H20.

Draw the standard normal curve and label the area is 95% below the value of X because it is 5% above this value.

Look up area=0.95 and get z= 1.645

Sub. Into z-formula
$z = \frac{x-\mu}{\sigma}$ $1.645 = \frac{x-100}{15}$ x=124.7

I. Practice Exam 1: Multiple Choice and Long Answer

Multiple choice:

I1. The answer is B. $b = r\frac{S_y}{S_x} = 0.975\,(\frac{250}{25})$

$$b = 9.75$$

$$a = \bar{y} - b\bar{x}$$

$$= 300 - 9.75(40) = -90$$

$$\hat{y} = -90 + 9.75x$$

I2.The answer is A. $\hat{y} = -90 + 9.75(20) = 105$

$$residual = y - \hat{y}$$

$$= 110 - 105$$

$$= 5$$

I3.The answer is C. $\frac{35+375}{490} = 0.837 \; or \; 83.7\%$

I4.The answer is D

I5. The answer is D. 0.8, 1.6, 2.8, [3.5], 4.2, 5.9, 8.2

Q1 = 1.6

Q3 = 5.9

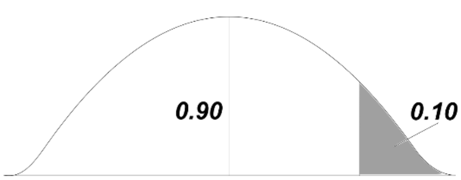
IQR = Q3 – Q1 = 5.9 – 1.6 = 4.3

Below Q1 – 1.5 IQR

= 1.6-1.5(4.3)

= -4.85

I6.The answer is A.



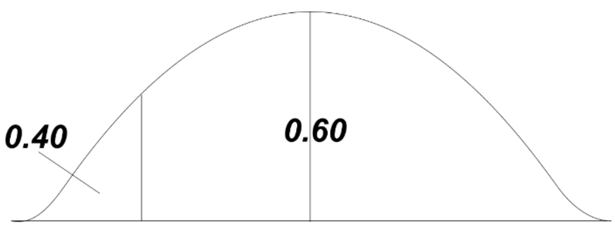
Look up Area 0.90 and get $z = 1.28$

$$z = \frac{x-\mu}{\sigma}$$

$$x = z\sigma + \mu$$

$$x = 1.28(10) + 75 = 87.8$$

I7. The answer is C.



$z = \frac{x-\mu}{\sigma}$

$-0.25 = \frac{x-67}{3.2}$

$x = -0.25(3.2) + 67 = 66.2$

I8.The answer is B. $\frac{25+30}{100} = 55\%$

I9.The answer is D.

Find Q3
IQR = Q3 – Q1
6000 = Q3 – 3000
Q3 = 9000

Find Maximum
range = max – min
15 000 = max – 2000
Max = 17 000

The mean is greater than the median, so it is right skewed.

I10. The answer is C. $z = \frac{x-\mu}{\sigma} = \frac{5-3.6}{1.5} = 0.93$

$\Pr(z > 0.93) = 1 - 0.8238$
$= 0.1762 \text{ or } 17.62\%$

I11. The answer is B. The numbers are the closest together, so the smallest standard deviation.

I12. The answer is A. multiply the standardized values together and find it divided by $(n-1)$
n-1=5-1=4

The standardized values are $(\frac{x_i-\bar{x}}{s_x}) and\ (\frac{y_i-\bar{y}}{s_y})$

$$[-1.2(-0.4) + 0.6(0.1) + (-0.25)(0.09) + (-1.04)(-2.2) + (1.25)(-1.61)] \div 4$$
$$= 0.793 \div 4 = 0.198$$
$$r = 0.198$$

I13. The answer is E).

I14. The answer is C) since graph A is more spread out it has a larger standard deviation. The means are equal.

I15. The answer is B). About 25% of the scores are above Q3=40.

I16. The answer is C).

Q1=20, Q3=40 and IQR=40-20=20

Outliers

below Q1 – 1.5 IQR
=20 – 1.5 (20)
= -10, so no outliers

Above Q3 + 1.5 IQR
= 40 + 1.5 (20)
= 70, so 80 is an outlier

I17. The answer is A $b = -0.23$

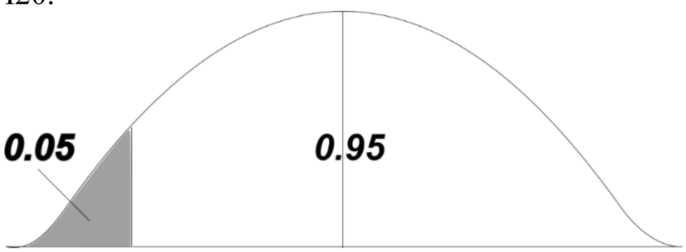
$$b = r\frac{s_y}{s_x}$$
$$-0.23 = r\ \frac{0.5}{2}$$
$$-0.23 = 0.25r$$
$$r = -0.92$$

I18. The answer is B). $z_1 = \frac{x_1-\mu}{\sigma}$ $z_1 = \frac{93-76}{5} = 3.4$

$$z_2 = \frac{88-72}{4.1} = 3.9$$

I19. The answer is A).

I20.



Look up 0.05 in body $z = -1.645$

$$z = \frac{x-\mu}{\sigma}$$

$$x = z\,\sigma + \mu$$

$$x = -1.645(15) + 110$$

$$x = 85.3$$

The answer is C).

Long Answer Questions

1.a) $b = r\frac{S_y}{S_x} = 0.98\left(\frac{200}{25}\right) = 7.84$

$$a = \bar{y} - b\bar{x}$$

$$= 500 - 7.84(38) = 202.08$$

$$\hat{y} = a + bx = 202.08 + 7.84x$$

$$\hat{y} = 202.08 + 7.84x$$

b) $\hat{y} = 202.08 + 7.8(100) = 986.08$

$$residual = y - \hat{y}$$

$$= 950 - 986.08$$

$$= -36.08$$

2. a)

Class	Returned	Non response	Total
First year	100	180	280
Second year	90	160	250
Third year	150	120	270
Fourth year	160	190	350
Total	500	650	1150

b) $\frac{160}{350 \leftarrow 4th\ year}$

$= 0.457\ or\ 45.7\%$

c) $\frac{90}{500} = 0.18\ \ or\ 18\%$

d) $\frac{180}{1150} = 0.157\ \ or\ 15.7\%$

3.a) IQR = Q3 – Q1

5000 = Q3 – 3000

Q3 = 8000

Range = max – min

14 000 = max – 2000

Max = $ 16 000

$\therefore$ *top* 25% *is from Q*3 *to* max $\therefore$ \$8000 *to* \$16 000

b) Outliers: below Q1 – 1.5 IQR

= 3000-1.5 (5000)

= - 4500, no outlier

Above Q3 + 1.5 IQR

= 8000 + 1.5(5000)

= \$15 500, so \$16,000 and \$17500 are outliers

4.a) $z = \frac{x-\mu}{\sigma}$

$\Pr(55 \leq x \leq 70)$

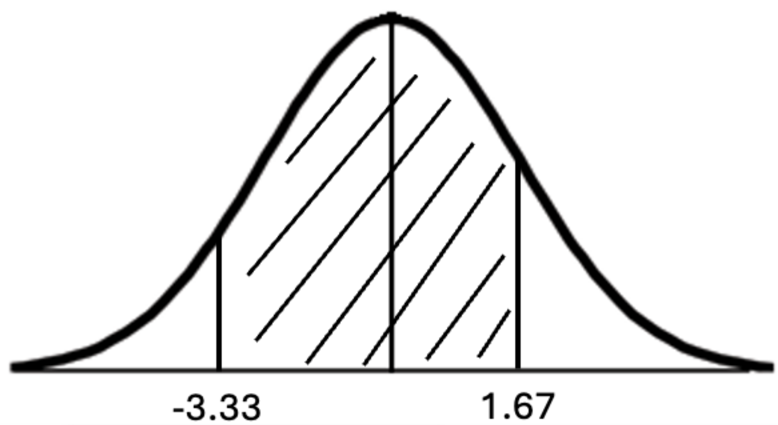
$\Pr(\frac{55-65}{3} < z < \frac{70-65}{3})$

$= \Pr(-3.33 < z < 1.67)$

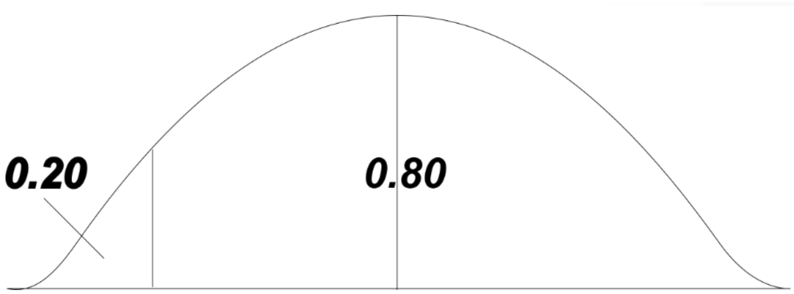
$= \Pr(z < 1.67) - \Pr(z < -3.33)$

$= 0.9525 - 0.0004$

$= 0.9521$



b)

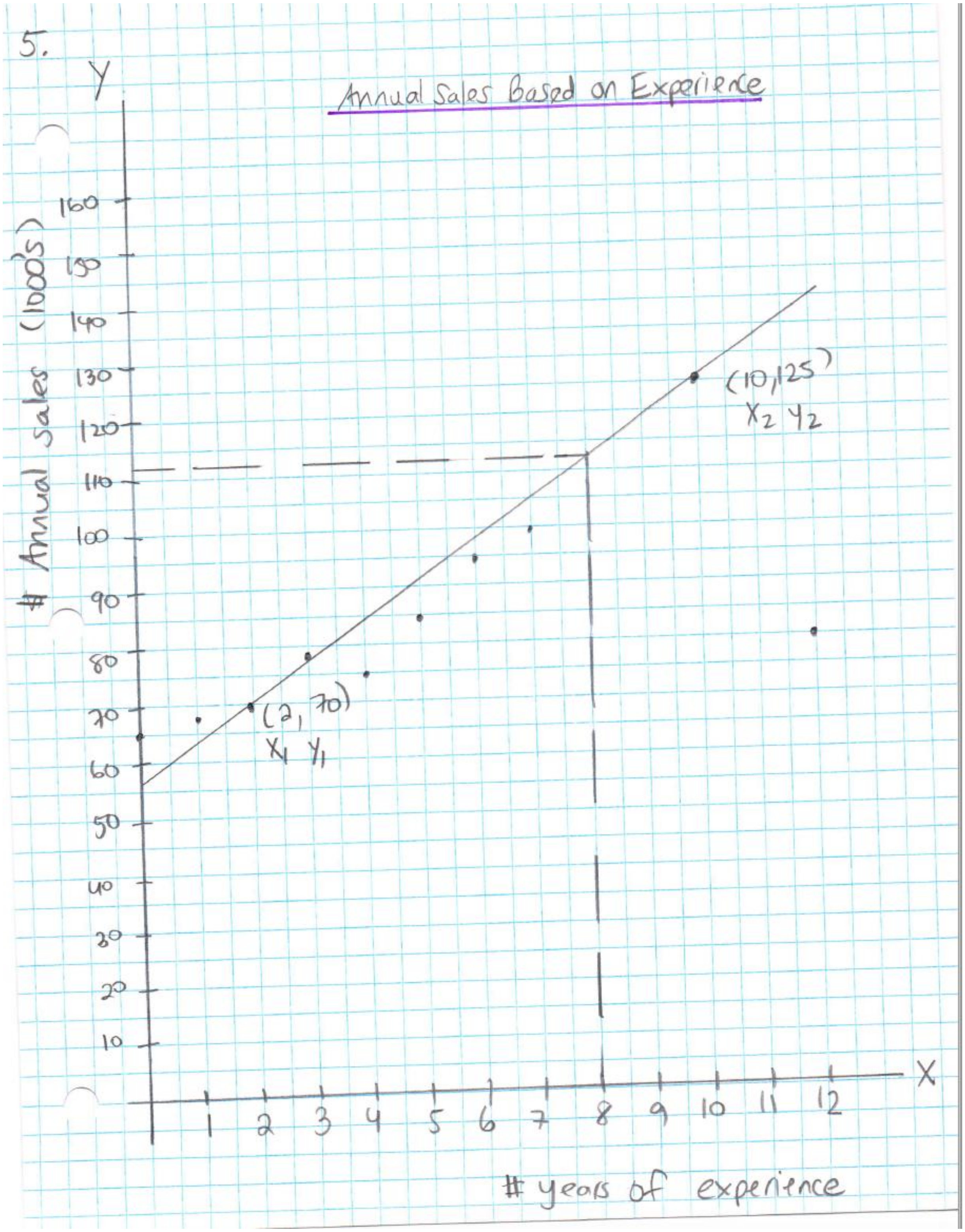


look up Area in body $z = -0.84$

$z = \frac{x-\mu}{\sigma}$

$x = z\,\sigma + \mu = -0.84(3) + 65 = 62.48$

Therefore, 80% are taller than 62 inches.


5.
Annual Sales Based on Experience
Y
Annual sales (1000's)
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
(10,125)
X_2 Y_2
(2, 70)
X_1 Y_1
X
1
2
3
4
5
6
7
8
9
10
11
12
years of experience

5.

b) $\text{slope} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{125 - 70}{10 - 2} = \frac{55}{8} = 6.875 = b$

y-int $\doteq$ 57 (from graph) $\therefore$ $a = 57$

$$\hat{y} = a + bx$$

$$\hat{y} = 57 + 6.875x$$

c) Yes, (12, 80) is an outlier.

d) $y = 120{,}000$ (in 1000's)

$\therefore$ subst $y = 120$ into the equation

$$\hat{y} = 57 + 6.875x$$

$$120 = 57 + 6.875x$$

$$120 - 57 = 6.875x$$

$$63 = 6.875x$$

$$x \doteq 9.164$$

$\therefore$ they have approx. 9.2 yr of experience.

e) See graph. Use interpolation (within our data)
with 8 yr of experience, we predict
their annual sales to be approx. \$112,000.

I.

6.a) 28, 32, 45, 65 median = $\frac{32+45}{2} = 38.5$

b) $mean = \frac{\sum x}{n} = \frac{28+32+45+65}{4} = 42.5$

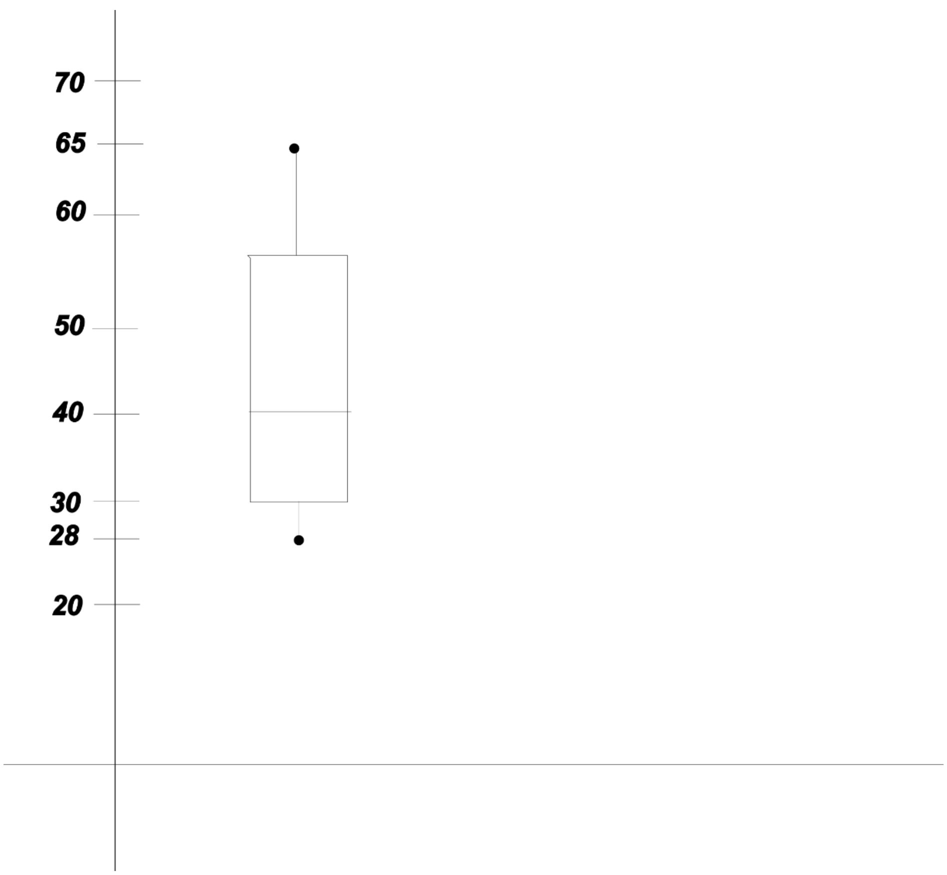
c) $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$

$$s = \sqrt{\frac{(45-42.5)^2 + (28-42.5)^2 + (32-42.5)^2 + (65-42.5)^2}{3}}$$

$= \sqrt{\frac{6.25+210.25+110.25+506.25}{3}}$

$= 16.7$

d)

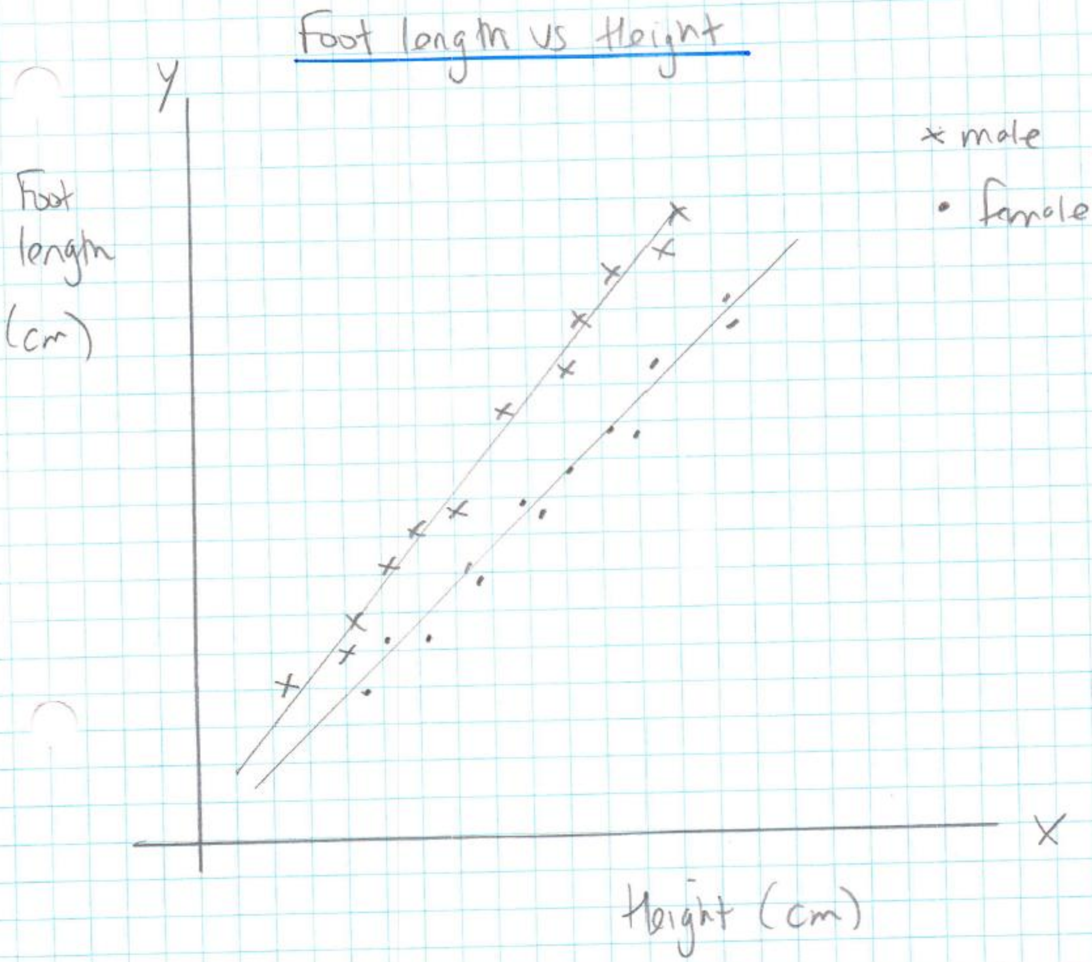


Q1 = $\frac{28+32}{2} = 30$ $Q3 = \frac{45+65}{2} = 55$

5 Number Summary: 28, 30, 38.5, 55, 65

7.

a)



gender – male / female categorical

foot length – quantitative

height – quantitative

as height ↑, foot length ↑

(positive correlation)

b) This data is categorical with the categories being different ranges of time of arrival to campus. The two graphs we talked about to use for categorical data are a pie chart and a bar graph. While a bar graph can be used for any set of categorical data, for a pie chart, the % must add to 100% exactly, so that we can measure out each "piece" of the pie as a percentage and then in degrees of the total circle.

A limitation is that if some people skip the question or select multiple categories, the total will not equal 100%, making a pie chart inappropriate, while a bar graph could still be used.

J. Practice Exam 2: Multiple Choice and Long Answer

Multiple choice:

J1. The mean would increase by 10, but the standard deviation would be the same as the numbers would have the same spread around the mean.

∴ The answer is B).

J2. The answer is A).

Since $\frac{6}{10} = 60\%\ \therefore true$

11, 12, 13, 20, 22, $|24, 33, 34, 42, 48$

Q1=13 Q3=34

IQR = 34-13 = 21 $\therefore true$

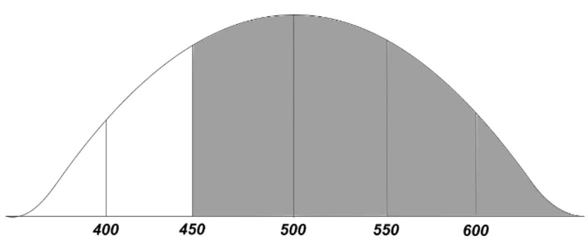
J3. $Z = \frac{x-\mu}{\sigma} = \frac{135-100}{2} = 2.92$

$\Pr(z < 2.92) = 0.9982$

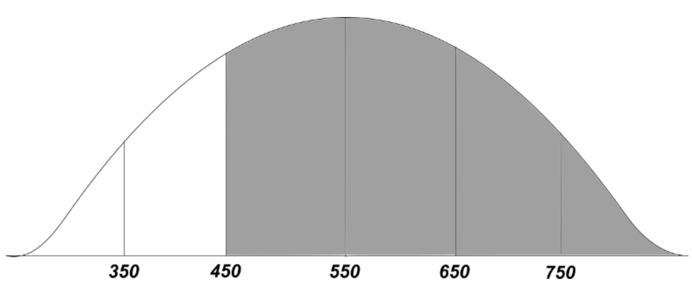
He scores higher than 99.82% of people

The answer is C).

J4. Boot 1



Boot 2



The answer is C).

J5. The answer is A).

Played $\frac{200+200}{1100} = 0.363$ Don't play $\frac{300+100}{1200} = 0.333$

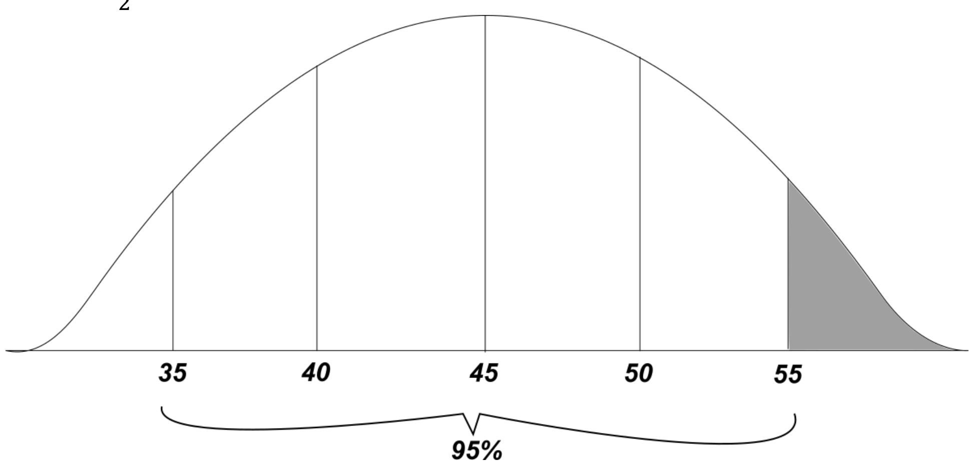
J6. $\frac{300}{300+200+200+400}$ =0.273

$\therefore 100 - 27.3\% = 73\%$

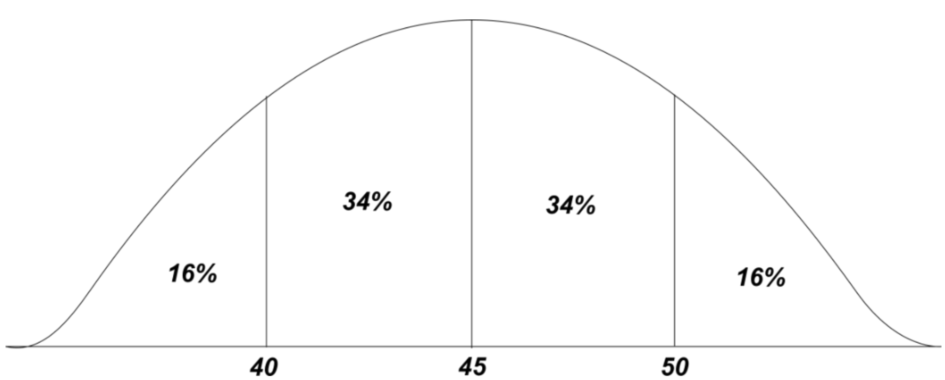
The answer is C).

J7. i) true – 50% lie above the mean

ii) true – $\frac{100-95}{2} = 2.5\%$



iii) true



The answer is A).

J8. The answer is D). We need to know the standard deviations to find b since $b = r\,\frac{s_y}{s_x}$
We only know that since r is positive, the slope would be positive.

J9. $\hat{y} = 20\,000 + 900(25) = \$42\,500$
Residual = y - $\hat{y}$ = \$45000 – 42500 = \$2500
The answer is B).

J10. The answer is C).

65000 = 20 000 + 900x
45000 = 900 x
x=50

J11. The answer is C). See regression and residuals section.

J12. $\bar{y} = \dfrac{6000}{10} = 600$

$b = r\,\frac{S_y}{S_x} = 0.95\left(\frac{300}{25}\right) = 11.4$

$a = \bar{y} - b\bar{x}$

$a = 600 - 11.4(40) = 144$

$\hat{y} = 144 + 11.4x$

$\hat{y} = 144 + 11.4(30) = 486$

The answer is A).

J13. $z = \dfrac{x-\mu}{\sigma} = \dfrac{20-15}{3} = 1.67$

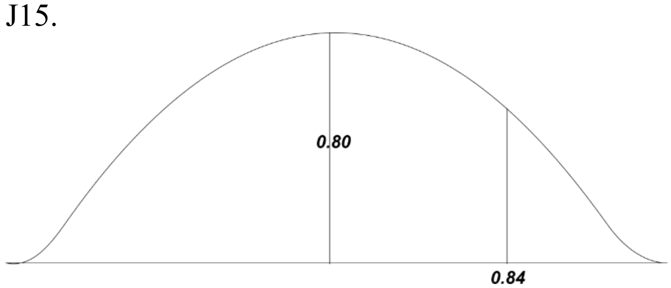
$\Pr(z > 1.67) = 1 - 0.9525$

$= 0.0475$

The answer is B).

J14. The answer is D).

J15.



Look up Area 0.80 in body $z = 0.84$

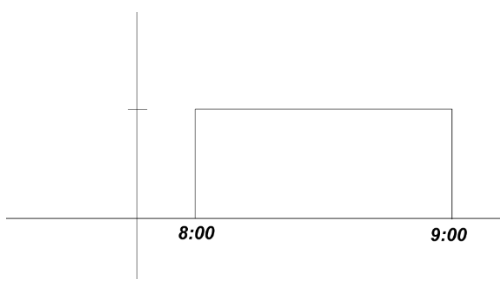
$z = \frac{x-\mu}{\sigma}$

$x = z\,\sigma + \mu$

$x = 0.84(25) + 100 = 121$

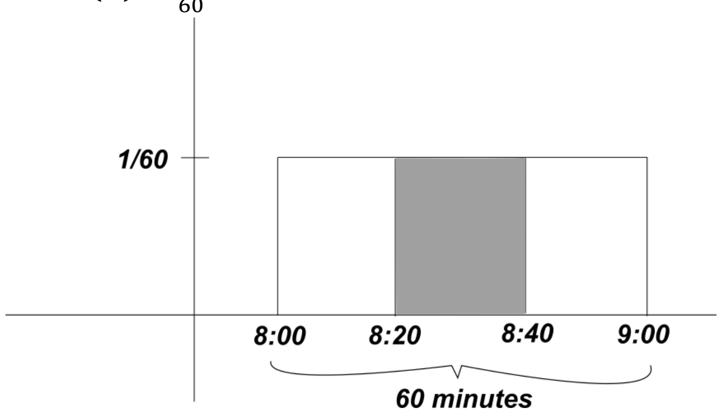
The answer is A).

J16.



$60(\quad) = 1$

$\therefore (\quad) = \frac{1}{60}$



$\Pr(8{:}20\ to\ 8{:}45) = L \times W$

$= 25 \times \frac{1}{60} = \frac{25}{60} = 0.42\ \ or\ 42\%$

The answer is C).

J17. $x = age$

$\hat{y} = \$ \; spent \; on \; repairs \; as \; x \uparrow, \hat{y} \uparrow$

$r^2 = 0.75^2 = 0.5625$ is explained by the model

$\therefore 1 - 0.5625 = 0.4375$ is not explained by the model

The answer is D).

J18. x, -0.8, 0.7, 0.9, 1.2, $\boxed{1.3}$, 2.5, 3.6, 4.2, 11.5, 12.8

Q1 = 0.7 Q3 = 4.2

IQR = Q3 – Q1 = 4.2 - 0.7 = 3.5

Outlier: Q1 – 1.5 IQR = 0.7-1.5(3.5)

= -4.55 below

The answer is B).

J19. Q1 occurs between 5th and 6th number

$\therefore 1 + 2 + 3 = 6 \quad \therefore 3$

The answer is B).

J20. $r = -\sqrt{0.95} = -0.975$ (negative since they are inversely related)

$$b = r\,\frac{s_y}{s_x} = -0.975\left(\frac{10}{100}\right) = -0.0975$$

The answer is C).

Long answer:

1.a) $r = \frac{1}{n-1}\sum \frac{(x_i - \bar{x})}{s_x}\frac{(y_i - \bar{y})}{s_y}$

$$r = \frac{1}{7}\frac{(-950)}{(13)(15)} = -0.696$$

b) $r^2 = (-0.696)^2 = 0.484 \;\; or \;\; 48.4\%$

c) $\bar{x} = \frac{50}{8} = 6.25 \qquad \bar{y} = \frac{60}{8} = 7.5$

$$b = r\,\frac{s_y}{s_x} = -0.696\left(\frac{15}{13}\right) = -0.803$$

$$a = \bar{y} - b\bar{x} = 7.5 - (-0.803)(6.25)$$

$$a = 7.5 + 5.01875 = 12.51875$$

$$\hat{y} = a + bx$$

$$\hat{y} = 12.52 - 0.803x$$

d) $\hat{y} = 12.52 - 0.803(10) = 4.49$

residual= $y - \hat{y} = 5.2 - 4.49 = 0.71$

2.a) $\sum x = 775$

$x = 775 - 50 - 78 - 82 - 90 - 100 - 150 - 200$

$x = 25$

b) median = $\frac{82+90}{2}$ = 86

Q1 = $\frac{50+78}{2} = 64 \qquad Q3 = \frac{100+150}{2} = 125$

IQR = Q3 – Q1 = 125 – 64 = 61

c) $s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1} = \frac{96\,533 - \frac{(775)^2}{8}}{7} = \frac{96\,533 - 75078.125}{7}$

$s^2 = 3064.982$

$s = 55.4$

$d) \frac{775 - 200}{7} = 82.1$

$e)\ below\ Q1 - 1.5\,IQR = 64 - 1.5(61) = -27.5$ none

$above\ \ Q3 + 1.5\,IQR = 125 + 1.5(61) = 216.5$ none

$\therefore no\ outliers$

3. a)

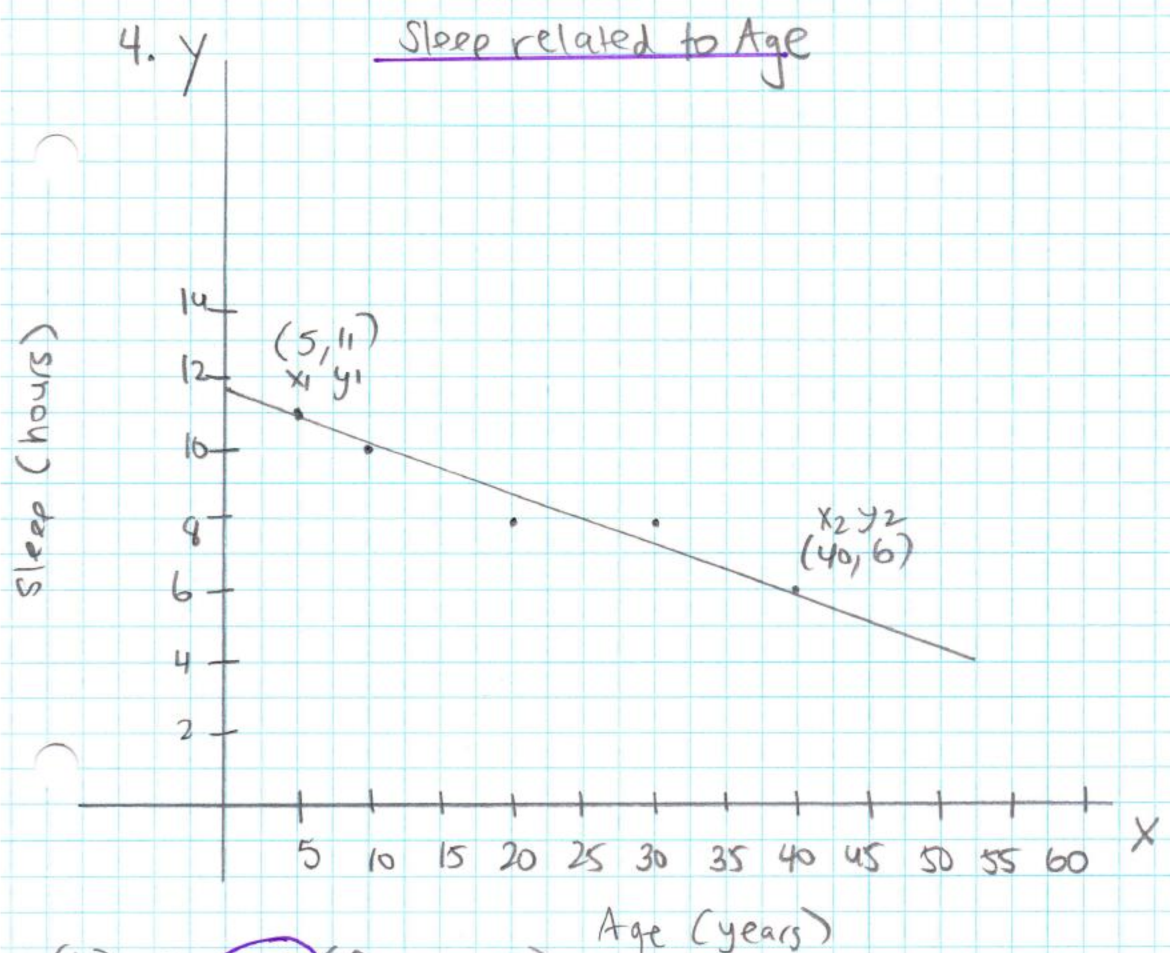
	Psychology	Sociology	Business	Total
Male	50	100	80	230
Female	60	120	70	250
Total	110	220	150	480

$b)\ \frac{50}{480} = 0.104\ \ or\ 10.4\%$

c) $\frac{120}{220} = 0.545\ \ or\ \ 54.5\%$

d) $\frac{70}{250} = 0.28\ \ or\ \ 28\%$

4.



(b) $a = 11.8$ (from graph)

$b = \frac{y_2 - y_1}{x_2 - x_1} = \frac{6 - 11}{40 - 5} = \frac{-5}{35} = -\frac{1}{7}$

$\hat{y} = a + bx$ $\therefore \hat{y} = 11.8 - \frac{1}{7}x$

(c) let $x = 25$ subst into the equation to find y

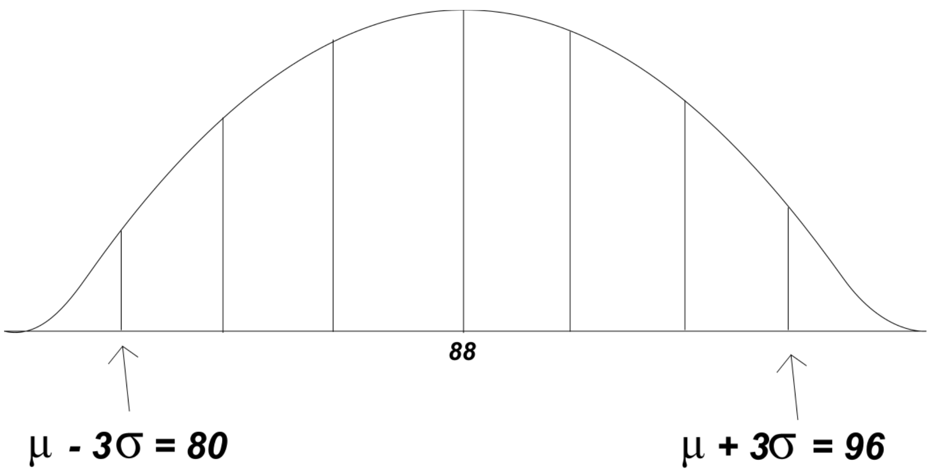
$\hat{y} = 11.8 - \frac{1}{7}x$

$\hat{y} = 11.8 - \frac{1}{7}(25)$

$\hat{y} = 11.8 - 3.57 = 8.2$

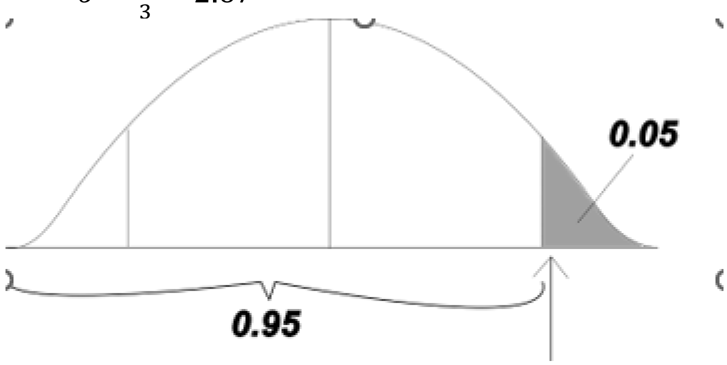
$\therefore$ A 25 year old is predicted to sleep 8.2 hr.

5.a)



$$\mu = \frac{80+96}{2} = 88$$

$$88 + 3\sigma = 96$$
$$3\sigma = 8$$
$$\sigma = \frac{8}{3} = 2.67$$

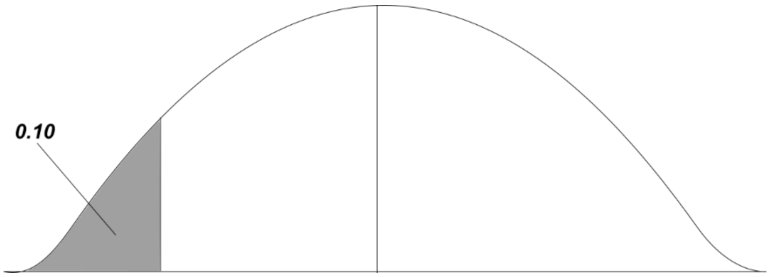


Look up the area below your line in the body of the Z table and you get 1.645

$x = z\sigma + \mu = (1.645)(2.67) + 88 = 92.4\%$

∴ 92.4 % is the cut-off average

b)



look up 0.10 in body $z = -1.28$

$z = \frac{x-\mu}{\sigma}$

$x = z\,\sigma + \mu = -1.28(2.67) + 88$

$x = 84.6$

∴ A mark of 85 would be above the lowest 10% of people applying

c)Pr $(82 < x < 92)$

Pr $(\frac{82-88}{2.67} < z < \frac{92-88}{2.67})$

$= \text{Pr}\,(-2.25 < z < 1.50)$

$= \text{Pr}(z < 1.5) - \text{Pr}\,(z < 2.25)$

$= 0.9332 - 0.0122$

$= 0.921$

∴ $200 \times 0.921 \cong 184$ *About* 184 *students scored between* 82 *and* 92.

6.

Age Groups	Fail/Success	Treatment A	Treatment B	Total
< 40	Fail	12	38	50
	Success	78	230	308
$40 +$	Fail	78	32	110
	Success	188	52	240

Combined the data:

Fail/Success	Treatment A	Treatment B	Total
Fail	90	70	160
Success	266	282	548

a) Treatment A

$< 40 \quad \text{Success} = \frac{78}{12+78} = \frac{78}{90} = \boxed{0.867}$

$40 + \quad \text{Success} = \frac{188}{78+188} = \frac{188}{266} = \boxed{0.707}$

Treatment B

$< 40 \quad \text{Success} = \frac{230}{38+230} = \frac{230}{268} = \boxed{0.858}$

$40 + \quad \text{Success} = \frac{52}{32+52} = \frac{52}{84} = \boxed{0.619}$

∴ the success rate is higher in both age groups for treatment *A* than treatment *B*

b) Combined Treatment A Treatment B

Success rate $= \frac{266}{90+266} = \frac{266}{356} = \boxed{0.747}$ $= \frac{282}{282+70} = \frac{282}{352} = \boxed{0.801}$

∴ when we combine the data, treatment *B* has a higher success rate than treatment *A*

c) Yes, this is an example of Simpson's Paradox because when the data was separated by age groups, Treatment A had a higher success rate for each age group. However, once the data was combined, Treatment B has a higher success rate. When the relationship reverses when the data is combined, this is what is referred to as Simpson's Paradox.

K. Methods of Sampling

p.154
Example. What is the target population and the sample?

Huron's administration wants to learn more about student reading preferences. They randomly assign each student a unique number and randomly generate 45 students to survey.

Target population: All Huron students

Sample: the 45 students that were selected randomly

p. 155
Example.
1. Observational study 2. Experiment

p. 158
Example. Suppose we conduct an observational study of the relationship between smoking during pregnancy and a child meeting their milestones in the first year of life. What are each of the variables?

Explanatory: smoking during pregnancy

Response: milestones met in the first year of life

Possible confounding variables: mother's education level, socioeconomic status, parent-child interactions

p.160

Example. A researcher wants to study the effects of a new fertilizer on 60 plants. The plants are: 10 tomato plants, 25 pepper plants, and 25 basil plants. These plants are known to grow at different rates.

a) Should you use a block design or completely randomized design and why?

You should use a block design since the type of plant will affect the rate of growth with and without fertilizer. If you mix all the plants together, you won't know if the fertilizer is causing differences in growth or if it is due to different types of plants.

b) If using blocks, what would be the blocks?

The blocks would be the different types of plant, i.e. tomato, pepper, and basil

p.161

Example. Summarize the Four Principles of Good Experimental Design:

1. Comparison groups- have a control group and a treatment group

2.Randomization- randomly assign subjects to groups

3. Blocking- group similar subjects prior to randomization

4. Replication- Use enough subjects to reduce variation occurring due to chance

p. 168

Example. A medication for arthritis has four dosage levels (5 mg, 10mg, 20 mg, 40 mg) and two delivery methods (pill, injection).

What are the factors? Dosage level, delivery method

How many factors are there? 2

How many treatments are there? 4(2) = 8

List a few of the possible treatments. E.g. 5 mg pill, 10 mg injection, 20 mg pill…

Example 1. The answer is (c).

Example 2. The answer is (d).

Example 3. The answer is (c).

Example 4. The answer is (d).

Example 5. The answer is (d).

Example 6. The answer is (b).

Example 7. The answer is (e).

Example 8. The answer is (e).

Example 9. The answer is (b).

Example 10. The answer is (c).

Example 11. The answer is (c).
Example 12. The answer is (b).

Example 13. The answer is (c).

Example 14. The answer is (b).

Example 15. The answer is (d).

Example 16. The answer is (a).

Example 17. The answer is (d).

Example 18. The answer is (d).

Example 19. The answer is (a).

Example 20. The answer is (d).

Example 21. The answer is (d).

Example 22. The answer is (a).

Example 23. The answer is (b). The two factors or explanatory variables are temperature and humidity.

Example 24. The answer is (c). We have 3 temperatures and 2 humidities, so 2(3)=6.

Example 25. The answer is (c).

Example 26. The answer is (c).

Example 27. The answer is (a).

Example 28.
11793 20495 05907 11384 44982 20751 27498 12009
Circle one number at a time and do so until you obtain three names

1,1,7,9...we would use 1, 7, 9 since we can't pick 1 twice...so, call, Chapman, Stamm and Wright
The answer is c).

Example 29.
81507 27102 56027 55892 33063 41842 81868 71035 09001
The first four to get the new medication are
8,1,5,7 since 0 doesn't represent a name

Then, we get **2**, 7, 1, 0, 2, 5, **6**, 0, 2, 7, 5, 5, 8, 9, 2, **3,** 3, 0, 6, 3**, 4.**.. The bolded ones are the ones we take since all others are repeats of the first four subjects who are already getting the medication
So, 2, 6, 3, 4 are the subjects to get the placebo...meaning, Chapman, Lovett, Dennis and Fitzgerald...note since there are only 8, we could just assume it was the four people we didn't get at the start, but since there could be 30 people, you need to know the method
The answer is (d).

Example 30.
A). The explanatory variable is the herbal tea. The answer is (b).
B). The confounding variable isn't a variable being studied, but any that will mess up your study and make the cause and effect difficult to prove. Since the elderly might be doing better from having extra visits and attention, their increased cheerfulness might be due to the company and have nothing to do with the tea.
The answer is (d).

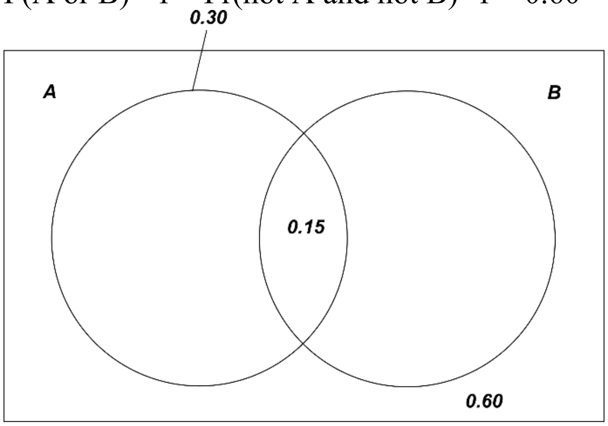
Example 31.
14 42 92 60 56 **31 42 48 03** 71 65 10 36 22 53 22 49 06
We would pick two numbers at a time, from left to right, until we get 5 numbers that are between 01 and 30
14, 42, 31, 48, 03...since we don't count 42 twice

The answer is (c).

L. Venn Diagrams

Example 1.

P(A or B)= 1 – Pr(not A and not B)=1 – 0.60 = 0.40



Pr(A or B)= Pr(A) + Pr(B) – Pr(A and B)
0.40 = 0.30 + 0.25 – Pr(A and B)
Pr(A and B) = 0.15
Pr(A and not B)= 0.30 – 0.15 = 0.15 (draw a Venn and subtract the middle)

Example 2.

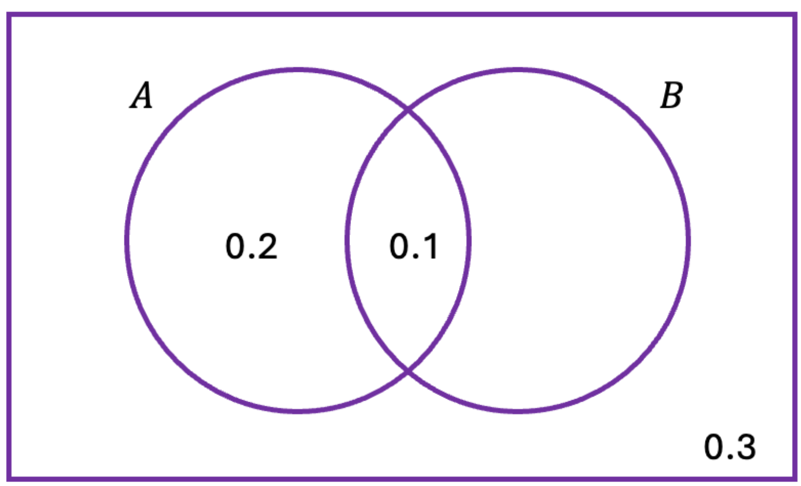
a) $\Pr(A \text{ or } B) = 1 - \Pr(\text{not A and not B}) = 1 - 0.3 = 0.7$

$\Pr(\text{A or B}) = \Pr(\text{A}) + \Pr(\text{B}) - \Pr(\text{A and B})$
$0.7=0.3+0.5-Pr(A \text{ and } \text{B})$
$Pr(A \text{ and } \text{B})=0.1$

b)

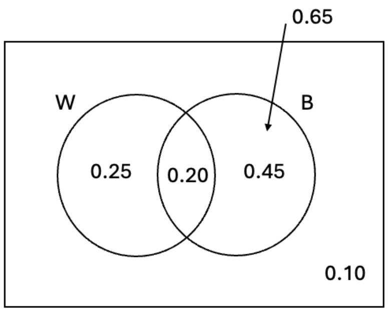
$\Pr(B^C)= 1 - \Pr(B) = 1 - 0.5 = 0.5$

$\Pr(A \text{ and } B^C) = \Pr(A) - \Pr(A \text{ and } B) = 0.3 - 0.1 = 0.2$
$\Pr(A \text{ or } B^C)= \Pr(A) + \Pr(B^C) - \Pr(A \text{ and } B^C)$
$= 0.3 + 0.5 - 0.2$
$= 0.6$



Example 3.

a) What percent of all degrees are earned by men?



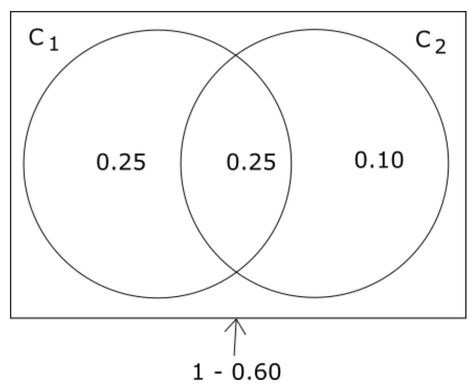
The total % earned by women is 45%
=100% - 45%=55%

b) What percent of all degrees are non-bachelor's degrees earned by men?

This is the outside of the Venn diagram, ie. 1 – 0.25 – 0.20 – 0.45 = 0.10 or 10%

Example 4.

(a) What is the probability that the company will win at least one of the two contracts?



Let $C_1 =$ "company wins first contract"

and $C_2 =$ "company wins second contract". Then

$$\Pr(C_1 or\, C_2) = \Pr(C_1) + \Pr(C_2) - \Pr(C_1 and\, C_2)$$

=0.50 + 0.35 – 0.25=0.60

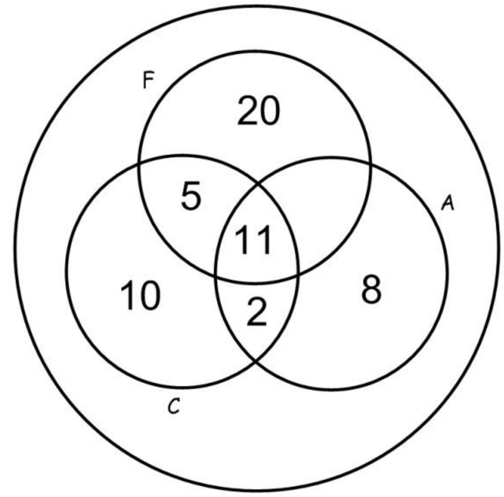
(b) What is the probability of winning the first contract but not the second?

Pr(C1 and not C2) = 0.25

(c) What is the probability of winning neither contract?

Pr(neither) = 1 – 0.60 = 0.40

Example 5.



a) From the Venn diagram, 11 students are taking all three courses.

b) Pr(only Finite) = 20/60 = 1/3

c) Pr(Calc and Algebra) = 11 +2 / 60 = 13/ 60

d) Pr(none of these three math classes) = 1 – (20+5+11+2+10+8)/ 60
= 1- 56/60
= 60/60 – 56/ 60
=4/60 or 2/30 or 1/15

M. Probability

Example 1.

BBB, BBG, BGB, BGG, GBB, GBG, GGB, GGG are all the possibilities.

Pr(exactly 2 girls)=3/8=0.375

Write down each outcome:

Sample Space, S

S= { FFF, FFM, FMF, FMM, MFF, MFM, MMF, MMM}

b) Event A=" first child is male"

A={MFF, MFM, MMM, MMF}

c) Event B= "at least one child is female"

B={ FFF, FFM, FMF, FMM, MFF, MFM, MMF}

d) Event C= "all three kids are male"

C= {MMM}

Example 2.

Pr(sum seven)=Pr{(1,6)(2,5)(3,4)(4,3)(5,2)(6,1)}=6/36=1/6

Example 3.

Pr(sum greater than 10)={(5,6)(6,5)(6,6)}=3/36=1/12

Example 4.

Sum	Win	Pr(W)
Less than or equal 4	2	6/36
5-10	-4	27/36
11,12	5	3/36

Pr(win $2)=Pr(roll 2 to 4)=1/36 + 2/36 + 3/36= 6/36= 0.17

Example 5.

P(A and B)= 1/36 only one outcome since it would be only {(5,2)}

P(AUB)=P(A or B)=P(A)+P(B)- P(A and B)

=6/36 + 6/36 - 1/36

=11/36

Example 6. a) Pr(red)=26/52

Pr(face card)=12/52

b)Pr(2 red)= $\frac{\binom{26}{2}}{\binom{52}{2}}$ or $\frac{26}{52} \times \frac{25}{51} = 0.245$

c) Pr(2 aces, with replacement)=$\frac{4}{52} \times \frac{4}{52} = \frac{1}{13}\left(\frac{1}{13}\right) = \frac{1}{169}$

Example 7.

Classical Probability Method

a) Write the sample space for rolling an ordinary die.

S= {1,2,3,4,5,6}

Pr(roll 1) = Pr(roll 2) = …= Pr(roll 6) = 1/6

b) What is the probability of rolling an odd number?

Pr(roll odd) = Pr(1,3,5) = 3/6 = ½

Non-Classical Probability Method

a) Write the sample space for rolling a die where the probability of rolling any odd number is twice the probability of rolling any even number.

Roll	Probability
1	2x
2	x
3	2x
4	x
5	2x
6	x

2x + x + 2x + x + 2x + x = 1 (probabilities add up to 1 in any experiment)

Solve for x:

9x = 1

x = 1/9

b) Pr (odd) = Pr (1,3,5) = 6x = 6 (1/9) = 6/9 or 2/3.

Example 8.

a) What is the sample space?

S= $\{(d_1, d_2, d_3)$ where $d_i \in \{0,1,2,3,4,5,6,7,8,9\}$

This means the sample space consists of all three-digit numbers where each digit is a number between 0 and 9 inclusive.

n(S) = 1000 since there are 1000 possible numbers

b) Let A= probability the number is less than 200
Find Pr(A).

A= { (0,0,0), (0,0,1), …(199)}

$$\Pr(A) = \frac{n(A)}{n(S)} = \frac{200}{1000} = 0.2\ or\ \frac{1}{5}$$

c) Let B= probability the number has three equal digits
Find Pr(B)

B={ (0,0,0), (1,1,1), … (9,9,9)} where n(B) = 10 possible numbers

$$\Pr(B) = \frac{n(B)}{n(S)} = \frac{10}{1000} = \frac{1}{100}\ or\ 0.01$$

Example 9. Pr(A and B)=Pr(A)Pr(B)= 0.2(0.4)=0.08

***Example 10.** Select a single card from a deck.

a) what is the probability it is a heart and a club?

A single card can't be both a heart and a club, so these are mutually exclusive events and Pr(both heart and club) = 0.

b) what is the probability it is either a heart or a club?

Pr(heart or club) = Pr(heart) + Pr(club) – Pr(both)

= Pr(heart) + Pr(club) - 0

= ¼ + ¼ = 2/4 = 1/2

c) what is the probability it is either a heart or an ace?

Pr(heart) + Pr(ace) – Pr(heart and ace)

= 13/52 + 4/52 – 1/52 since there is only 1 ace of hearts

=16/52 = 0.308

d) A= draw a diamond

A^c= not drawing a diamond, i.e. drawing a club, heart or a spade

Example 11.

Pr(B)=0.3 and Pr(A)=0.5, Pr(B or C)=0.7

a) Pr(B or C)=Pr(B) + Pr(C) - Pr(B and C)

B and C are independent, so Pr(B and C)=Pr(B)xPr(C)

So,

Pr(B or C)=Pr(B) + Pr(C)- Pr(B)xPr(C

0.7=0.3 + Pr(C) - 0.3Pr(C)

0.4=0.7Pr(C)....since 1Pr(C) - 0.3Pr(C)=0.7Pr(C)
Pr(C)=4/7

b) Pr(A or B)=Pr(A) + Pr(B) - 0 since A and B are mutually exclusive

=0.5+0.3=0.8

Example 12.

The answer is d). If they are mutually exclusive then Pr(A and B)=0...if they were to be independent as well, then Pr(A)xPr(B)=0 and this is impossible since we are told that the probabilities of A and B are non-zero.

Therefore, they would have to be DEPENDENT.

Example 13.

$\Pr(A\ and\ B) = \Pr(A) \times \Pr(B) = 0.4(0.3) = 0.12 \neq 0$ so, a) is true

since A, B are independent we multiply

Pr(A or B)=Pr(A) + Pr(B) - 0.12

=0.4 + 0.3 - 0.12

=0.58

So, b) is true

To check c)...$\Pr(not\ A\ and\ not\ B) = 1 - \Pr(AorB) = 1 - 0.58 = 0.42$ so, c) is true.

The answer is e).

Example 14.

Pr(H)=0.3 $\quad$ Pr(TTT)=(0.7)(0.7)(0.7)=0.7^3

Pr(T)=0.7

Pr(at least 2 T)=Pr(TTT) + Pr(TTH) + Pr(THT) + Pr(HTT)

$0.7^3+0.7^2(0.3)+0.7(0.3)(0.7)+(0.3)(0.7)^2=0.784$

Example 15.

Pr(T)=0.9

Pr(E)=0.8

It can be solved by Tina, Eddie or both of them

Pr(solved)=Pr(T and E)+Pr(T and not E) +Pr(E and not T)

=0.9(0.8) + (0.9)(0.2)+0.1(0.8)=0.98

or do 1-Pr(not solved)= 1 - 0.1(0.2)=1-0.02 = 0.98

Example 16. R and C are independent...

$$\Pr(not\ R\ and\ not\ C) = \Pr(not\ R)\ xPr(not\ C) = 0.2(0.3) = 0.06$$

Example 17. Pr(at least 1 fish)= 1 - Pr(no fish)

$=1-(0.4)(0.4)(0.4)(0.4) = 1-(0.6)^4 = 0.8704$

(independent since catching a fish doesn't affect the chance on each trial)

Example 18.

a) They are independent because the first flip being tails won't affect the second flip.

b) They are independent, since "ace" and "spades" don't affect each other. One is the type of suit and one is the denomination...i.e. we can get an ace of spades

c) These are disjoint, since one card can't be both a spade and a heart, i.e. prob. of both = 0

Example 19.

Pr(A or B) =Pr(A) + Pr(B) – Pr(A and B)= 0.3 + 0.2=0.5

Pr(A or C) =Pr(A) + Pr(C) – Pr(A and C)= 0.3+0.4=0.7

Pr(A or B or C)= 0.3+0.3+045=0.9

The answer is d).

b) Pr(A or not B) = Pr(A) + Pr(not B) – Pr(A and not B)

=0.3 + (1-0.2) – Pr(A) since all of A is not in B as they are disjoint

=0.3 + 0.8 – 0.3 = 0.8

N. Conditional Probability

Example 1.

$$\Pr(E/F) = \frac{\Pr(E \text{ and } F)}{\Pr(F)}$$

$$0.375 = \frac{0.30}{\Pr(F)}$$

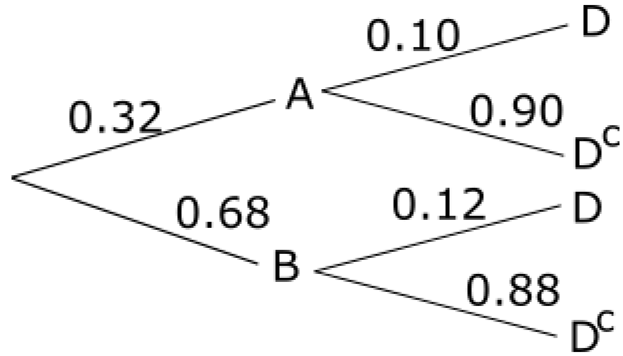
0.375 Pr(F)= 0.30
Pr(F) = 0.30/0.375=0.80

Example 2.

$$\Pr(F/E) = \frac{\Pr(F \text{ and } E)}{\Pr(E)} = \frac{0.20}{0.40} = 0.5$$

Example 3.

$$\Pr(B/D) = \frac{\Pr(B \text{ and } D)}{\Pr(D)} = \frac{(0.68)(0.12)}{(0.32)(0.10)+(0.68)(0.12)} = 0.72$$



Example 4.
a) Pr(S)=70/200=0.35

b) Pr(S/M)=30/100=0.30

c) Pr(F/S)=40/70=0.571

Example 5.

Pr(S)=0.40

Pr(D)=0.55
Pr(D/S)=0.75

$$\Pr(D/S) = \frac{\Pr(D \cap S)}{\Pr(S)}$$

Pr(D and S)=(0.75)(0.40)=0.3

Example 6.

Pr(B)=0.57
Pr(D)=0.82

Pr(B and D)=0.45

Pr(D/B)=$\frac{Pr(D\ and\ B)}{Pr(B)} = \frac{0.45}{0.57} = 0.79$

Example 7.

Pr(A)=1/2 and Pr(B)= 3/8=0.375

Pr(A and B)=0.20

Pr(A/B)=$\frac{0.20}{0.375} = 0.53$

Example 8.

Pr(A or B)=Pr(A) + Pr(B) - Pr(A and B)

0.9 = 0.35 + 0.75 - Pr(A and B)

Pr(A and B)=0.20

A) true

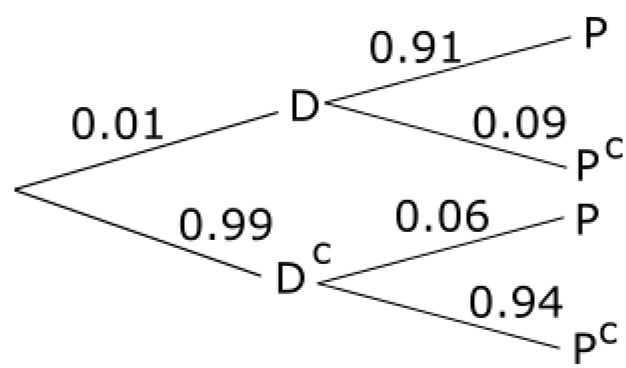
B) Pr (A/B)=$\frac{\Pr(A\ and\ B)}{\Pr(B)} = \frac{0.20}{0.75} = 0.27$

C) Pr(B/A)= $\frac{\Pr(A\ and\ B)}{\Pr(A)} = \frac{0.20}{0.35} = 0.57$

So, the answer is D).

Example 9.

D= has disease
D^C= doesn't have disease
P= tests positive
P^C=tests negative

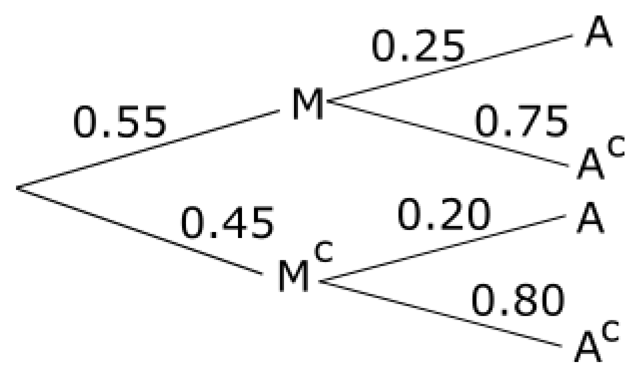


a) Pr(correct) = Pr(D and P) + Pr(D^C and P^C)
=0.01(0.91) + 0.99(0.94)
=0.9397

b) Pr (D/P)=$\frac{\Pr(D\ and\ P)}{\Pr(P)} = \frac{0.01(0.91)}{0.01(0.91)+0.99(0.06)} = 0.13$

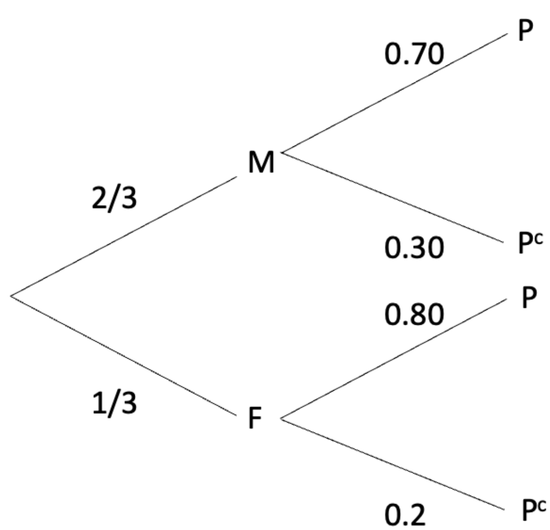
Example 10.

Pr(M/A)=$\frac{\Pr(M\ and\ A)}{\Pr\ (A)} = \frac{0.55(0.25)}{0.55(0.25)+0.45(0.20)} = 0.604\ or\ 60.4\%$



Example 11.

$\Pr(F/P^C)$
$= \frac{\Pr(F \cap P^C)}{\Pr(P^C)} = \frac{1/3(0.20)}{2/3(0.30)+1/3(0.20)}$=0.25



Example 12.

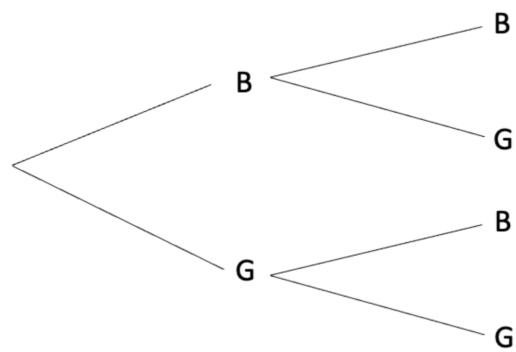
Reduced $S \cdot S = \{1\text{st roll } 2\} = \{(2,1)(2,2)(2,3)(2,4)(2,5)(2,6)\}$

Pr (sum > 6/1st roll is a 2)

$= \frac{2}{6} \begin{matrix} \leftarrow \text{sum} > 6 \\ \leftarrow \text{1st roll a 2} \end{matrix}$

$= \frac{1}{3}$

Example 13.



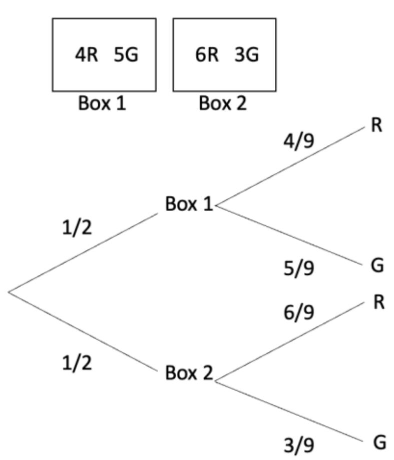
Outcomes BB, BG, GB, GG
Reduced $S \cdot S =$
{BB, BG, GB} since they have at least 1 boy

∴ Pr (2 boys/at least 1 boy)

$= \frac{1}{3} \begin{matrix} \leftarrow \text{2 boys (BB)} \\ \leftarrow \text{3 outcomes} \end{matrix}$

Example 14.

a)



$\Pr(G) = \Pr(Box1 \cap G) + \Pr(Box2 \cap G)$
$= \frac{1}{2}\left(\frac{5}{9}\right) + \frac{1}{2}\left(\frac{3}{9}\right) = \frac{5+3}{18} = \frac{8}{18} = \frac{4}{9}$

b)

$$\Pr(Box\ 2\ /G) = \Pr\frac{(\text{Box2} \cap \text{G})}{\Pr(G)}$$

$$= \frac{\frac{1}{2}\left(\frac{3}{9}\right)}{\frac{4}{9}} = \frac{3}{18} \times \frac{9}{4} = \frac{1}{6} \times \frac{9}{4} = \frac{9}{24} = \frac{3}{8}$$

Example 15.

If you flip a coin, and I roll a die, we can't possibly affect each other!

So, the probability I flip heads and you roll a 1 is:

Pr(H) × Pr(roll 1)

$= \frac{1}{2} \times \frac{1}{6} = \frac{1}{12}$

Example 16.

$\Pr(H) = \Pr(I\ and\ P\ and\ H)$

= $\Pr(I)\Pr(P\ /I)$ Pr $(H/$ I and P)

$= (0.5)(0.3)(0.2)$=0.03

*To find "AND", we multiply through our probability tree!!!!

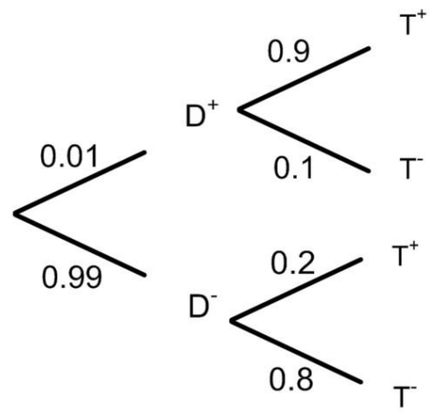
Bayes' Theorem

Example 17.

$$\Pr(B/A) = \frac{\Pr(A/B)\Pr(B)}{\Pr(A)} = \frac{\frac{1}{4}\left(\frac{1}{2}\right)}{\frac{1}{3}} = \frac{\frac{1}{8}}{\frac{1}{3}} = \frac{1}{8}\times\frac{3}{1} = \frac{3}{8}$$

Example 18.

$$\Pr(T^+/D^+) = sensitivity = 0.9$$
$$\Pr(T^-/D^-) = specificity = 0.8$$



$$\Pr(D^-/T^-) = \frac{\Pr(D^- and\ T^-)}{\Pr(T^-)} = \frac{0.8(0.99)}{0.01(0.1)+0.99(0.8)} = 0.999$$

$\uparrow$ *given*

Example 19.

	prob
Muffin	$3x$
Cookie	$6x$
Tart	x
	$10x$

← $\frac{1}{3}$ muffins

$$3x + 6x + x = 1$$
$$10x = 1$$
$$x = \frac{1}{10}$$

$$\Pr(muffin/C) = \frac{\Pr(muffin \cap C)}{\Pr(C)} = \frac{\frac{3}{10}(0.75)}{\frac{6}{10}(0.90)+\frac{1}{10}(0.25)+\frac{3}{10}(0.75)} = 0.285$$

O. Random Variables

Example 1.

x	Pr (x)	F(x)
1	1/5=2/10	2/10
2	1/10	3/10
3	7/10	10/10 = 1

We can find the missing probability for Pr(X=3) by remembering that all of the probabilities must add up to 1.

i.e. $\Pr(X = 3) = 1 - \frac{1}{5} - \frac{1}{10} = \frac{10}{10} - \frac{2}{10} - \frac{1}{10} = \frac{7}{10}$

Example 2.

Pr(8<x<15)= Area rectangle = (length)(width)==(15-8)(1/10)=(7)(1/10)=0.70

Example 3.

a) This is a continuous random variable, since it is a uniform distribution, it is the area under a rectangle that we are finding

b) This is another continuous random variable, since it is a normal distribution, so it is the area under the standard normal curve we are finding

c) This is a discrete random variable as there is simply a countable number of values of x, from 0 to 5.

d) This is a discrete random variable as it is a countable number of people who visited Swiss Chalet during the lunch period

Example 4.

Pr(4<x<15) = LxW= (15-4) (1/20) = 11/20 = 0.55

Example 5. Pr(X<2000) = bxh/2 = (1000)(0.0002)/2 = 0.10

Example 6.

To find the value of a, find the area of each of the shapes and set it equal to 1, since the total area is the same as the total probability

A= $\frac{(2)\left(\frac{1}{2}a\right)}{2} + (6)\frac{1}{2}a = 1$

$$1 = \frac{1}{2}a + 3a$$

$$1 = \frac{7}{2}a$$

$$a = \frac{2}{7}$$

P. Final Exam Questions on Sections J to O

P1. A) Stratified sample

B) d) ABC radio status

P2. (a) TRUE. Yes, because the people getting the vitamins don't know whether they are getting the real thing or a fake pill, a placebo.

(b) FALSE. A double-blind study is when both the people receiving the treatment and the person handing out the pills BOTH don't know who is getting the real thing.

(c) TRUE. Yes, a placebo is a sugar pill and it is used in this experiment.

P3. (a) we cannot separate their effects on a response variable.

P4. (a) a stratified sample. The members are split into groups and then a random sample is taken from each group. This is stratified sampling.

P5. (c) a multi-stage design. The groups are separated into groups, like they would be in stratified, but here the groups are based on poor or wealthy communities. Then three houses are randomly selected in each neighbourhood.

P6. a) SRS

b) Stratified sampling

c) Systematic sampling.

P7. Choose a SRS of 3 people from the following students:

Abby 00
Brooke 01
Cole 02
Dennis 03
Edward 04
Frankie 05
Grace 06
Harrison 07
Kelly 08
Maureen 09

Use line 130 from Table B.

69051 64817 87174 09517 84534 06489 87201 97245 05007 16632 81194 14873
04197 85576 45195

Go through the numbers from left to right and look at two-digit numbers and circle any that are in the set "01, 02,...10".

The numbers you get are 05, 00, 04.

So, we choose Frankie, Abby and Edward
NOTE: you can't pick 05 twice

P8. 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20
Line 101 is:
19223 95034 05756 28713 96409 12531...

We would choose 19, 05, and 13

P9. $\Pr(A) = 1 - \Pr(O) - \Pr(B) - \Pr(AB) = 1 - 0.50 - 0.20 - 0.05 = 0.25$. The answer is (c).

P10. Pr(both aces)= $\frac{4}{52} \times \frac{3}{51}$=0.00452
or if you know the choose formula from 1228, $\frac{\binom{4}{2}}{\binom{52}{2}}$

P11. $\Pr(A\ and\ B) = \Pr(A) \times \Pr(B) = 0.2(0.3) = 0.06 \neq 0$ since A, B are independent
So, a) is true
Pr(A or B)=Pr(A) + Pr(B) - 0.06
=0.2 + 0.3 - 0.06
=0.44

b) is true
To check c)...$\Pr(not\ A\ and\ not\ B) = 1 - \Pr(A\ or\ B) = 1 - 0.44 = 0.56$
The answer is d). only a) and b) are true.

P12. Let E denote the event that a mosquito was a carrier of the virus. Then E^C denotes the event that the mosquito was not a carrier of the virus. Since each mosquito has a 90% of not being a carrier of the virus,

$$Pr(\ E^C) = (0.90)^4 = 0.6561.$$

Therefore $Pr(E) = 1 - Pr(E^C) = 1 - (0.90)^4 = 0.3439 = 34.39\%.$

P13. The probabilities of drawing 1 red ball, 1 green ball, or 1 yellow ball are

$$\Pr(R) = \frac{5}{10}, \qquad \Pr(G) = \frac{3}{10}, \qquad \Pr(Y) = \frac{2}{10},$$

respectively.
The probabilities of drawing 2 red balls, 2 green balls, or 2 yellow balls are

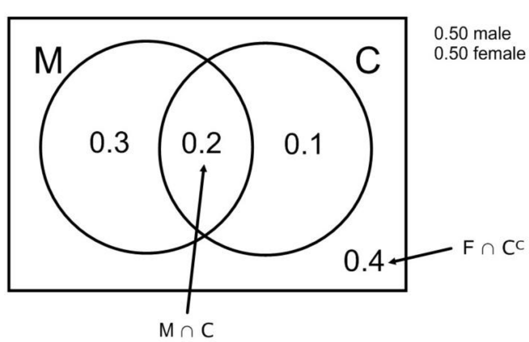
$$\Pr(RR) = \left(\frac{5}{10}\right)^2, \qquad \Pr(GG) = \left(\frac{3}{10}\right)^2, \qquad \Pr(YY) = \left(\frac{2}{10}\right)^2,$$

respectively.
The probability of drawing 2 balls of the same colour is therefore

$$\Pr(RRorGGorYY) = \left(\frac{5}{10}\right)^2 + \left(\frac{3}{10}\right)^2 + \left(\frac{2}{10}\right)^2 = 0.25 + 0.09 + 0.04 = 0.38.$$

P14. If 30% have a college degree and 20% of men have a college degree, then 10% of the women have a college degree

Pr(female and college degree)= 0.10....female without college would be 0.4, if they asked!



P15. The probability of *not* catching a fish each time you cast your line is $1-\frac{1}{4}=\frac{3}{4}$.
The probability of *not* catching a fish on the first two attempts is $(\frac{3}{4})^2=\frac{9}{16}$.
The probability of catching at least one fish within the first two attempts is thus $1-\frac{9}{16}=\frac{7}{16}$.
The answer is (b).

P16. Pr(F)=0.40 and Pr(N)=0.30, Pr(F and N)=0.20

Pr(F or N)= Pr(F) + P(N) – Pr(F and N)=0.40+0.30-0.20=0.50

P17.
a) Pr(40-49)=(10+15+50+70)/400
=145/400=0.3625

b) 50/400

c)145+55/400= 200/400=0.5

d) 15+10/400= 25/400=0.0625

e) 60+30/400=90/400=0.225

P18.

BBB, BBG, BGB, BGG, GBB, GBG, GGB, GGG are all the possibilities.

Pr(exactly 2 girls)=3/8=0.375

P19.
Pr(sum greater than 10)=Pr(sum 11 or 12)=Pr{(5,6)(6,5)(6,6)}=3/36=1/12

P20.
a) Yes, they are disjoint as you can't be underweight and obese at the same time...you can only belong to one category

b) Pr(D)= 1 - 0.02 - 0.39 - 0.35=0.24

P21. Pr(red)=26/52
Pr(face card)=12/52

P22. Since A and B are independent...
Pr(A and B)=Pr(A)Pr(B)= 0.2(0.5)=0.10

P23. The following table shows the distribution of blood types in 100 people.

	O	A	B	AB	Total
Rh Positive	39	35	8	4	86
Rh Negative	6	5	2	1	14
Total	45	40	10	5	100

a) If one person is randomly selected, find the probability they have AB blood type

5/100=0.05

b) If one person is randomly selected, find the probability they are O blood type or Rh negative.
Pr(O or Rh^-)= Pr(O) + Pr(Rh^-) - Pr(O and Rh^-)
=45/100 + 14/100 - 6/100
=53/100
=0.53

c) If one person is randomly selected, find the probability they are A blood type and Rh positive.

35/100=0.35

P24.

$\frac{40}{100} \times \frac{39}{99}$=0.158

P25. Suppose events A, B and C are all events in a sample space. You are given that A and B are mutually exclusive and B and C are independent, where Pr(B)=0.1, Pr(A)=0.4 and $\Pr(B \cup C) = 0.6$.

Find each of the following:

Pr (C)

$\Pr(B\ or\ C) = \Pr(B) + \Pr(C) - \Pr(B) \times \Pr(C)\ since\ B, C\ are\ independent$
0.6=0.1 + Pr(C) - 0.1Pr(C)
0.5 = 1Pr(C) – 0.1 Pr(C)
Pr(C)=0.5/0.9=5/9

P26. Pr $(A\ or\ B)$=Pr(A) + Pr(B) - 0 since they are mutually exclusive
=0.4+0.1=0.5

P27. P(A and B)= 1/36 only one outcome since it would be only {(6,1)}
P(A or B)=P(A or B)=P(A)+P(B)- P(A and B)
=6/36 + 6/36 - 1/36
=11/36

P28.
$\frac{2}{10} \times \frac{1}{10} \times \frac{2}{10} = \frac{4}{1000} = 0.004$

P29.
$\Pr(at\ least\ 1\ catch) = 1 - \Pr\ (none\ catch)$
$= 1 - (0.70)(0.60)(0.90)$..we can multiply since they are independent
$= 0.622\ \ or\ \ 62.2\%$

P30.
$June = 30\ days \qquad July = 31\ days$
$\therefore \frac{30+31}{365} = 0.17\ \ or\ \ 17\%$

P31.

	Snow	No snow	Total
Forecast snow	76	136	212
Forecast no snow	24	264	288
Total	100	400	500

$\therefore correct = \frac{76+264}{500} = \frac{340}{500} = 0.68\ \ or\ 68\%$

P32.
$C = 1 - 0.45 - 0.40 - 0.04$
$\therefore C = 0.11$
$\Pr(same) = \Pr(0,0) + \Pr(A, A) + \Pr(B, B) + \Pr\ (AB, AB)$
$= 0.45^2 + 0.40^2 + 0.11^2 + 0.04^2$
$= 0.3762\ \ or\ \ 37.62\%$

P33. Area = LxW= (7.5 – 2.5)(1/8)= 5/8

P34. Circle one number at a time and then select the restaurants that correspond to each number.

In the random list, we would use 6, 9, 0 and 4.

So, we would survey Archie's, Taco Bell, Wendy's and Jack Astor's.

P35. Area =1 - bxh/2= 1 – (450-200)(0.002)/2= 1 – 0.25= 0.75

P36. As soon as there are more than 10 restaurants, we would need to use two digits to label them

01Wendy's
02Tim Hortons
03Swiss Chalet
04Burger King
05Jack Astors
06McDonald's
07Archie's
08East Side Marios
09Montanas
10Taco Bell
11Harveys
12Pizza Hut
13Red Lobster

If we use the same random list of numbers, we need to look for two digit numbers that APPEAR in our list. A number such as the first two-digit number "69" doesn't apply because we don't have any restaurant labeled 69.

69043 81235 90721 30174 97245

69, ***04***, 38, ***12***, 35, 90, 72, ***13, 01***, 74, 97, 24,

Therefore, in this lists 04, 12, 13 and 01 are the first four numbers that actually represent restaurants. So, we would call Burger King, Pizza Hut, Red Lobster and Wendy's.

P37. Subjects must all have the same number of digits...they will be 01, 02, ..., 16

Circle two numbers at a time, without skipping any, until you find four people

81**05**7 271**02** 56**02**7 55892 33**06**3 41842 81868 7**10**35 43367

We got 02 twice, but you can't count the same person twice, so we get 05, 02, 06 and 10

P38. $\Pr(A/B) = \frac{\Pr\,(A\ and\ B)}{\Pr\,(B)}$

$\frac{2}{5} = \frac{\Pr\,(AandB)}{\frac{1}{2}}$

$\therefore \Pr(Aand\ B) = \frac{2}{5} \times \frac{1}{2} = \frac{1}{5}$

$\Pr(B/A) = \frac{\Pr\,(A\ and\ B)}{\Pr\,(A)} = \frac{\frac{1}{5}}{\frac{1}{3}} = \frac{1}{5} \times \frac{3}{1} = \frac{3}{5}$

P39.

a) $\Pr(not\ B) = 0.70(0.40) + 0.30(0.70)$
$= 0.28 + 0.21$
$= 0.49$

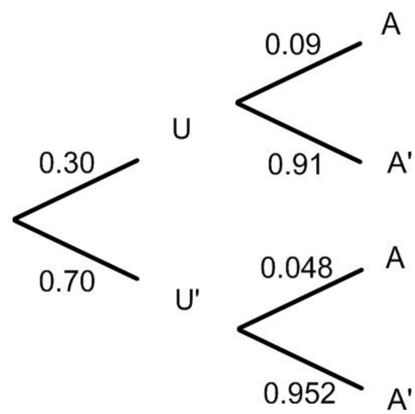
b) $\Pr(M/B) = \frac{\Pr\ (M\ and\ B)}{\Pr\ (B)} = \frac{0.30(0.30)}{1-0.49} = 0.176$

P40.
$U = under\ 25$
$P(U) = 0.30\ \ under\ 25$

$$\Pr(U/A) = \frac{\Pr\ (U\ and\ A)}{\Pr\ (A)}$$

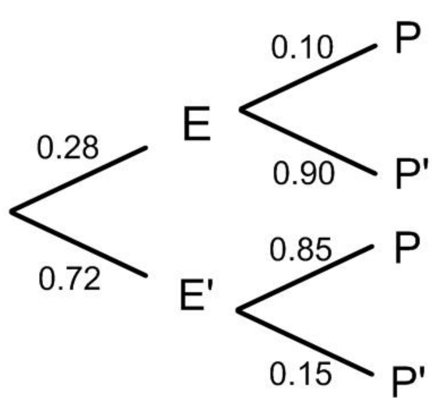
$$= \frac{0.3\times 0.09}{0.3\times 0.09 + 0.7\times 0.048}$$

$$= 0.446$$



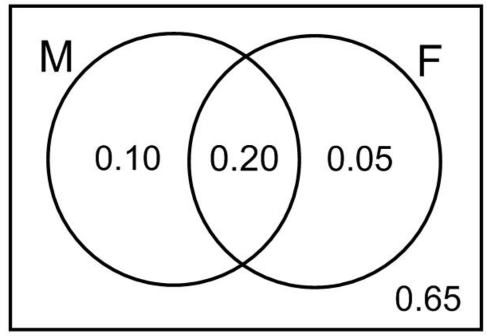
P41. Let E= emits excessive pollutants and let P= passes the test for emissions

$$\Pr\ (E/notP) = \frac{\Pr(E\ and\ not\ P)}{\Pr\ (not\ P)} = \frac{0.28(0.90)}{0.28(0.90)+0.72(0.15)} = 0.70$$



P42. Let M= got an A on the midterm and let F= got an A on the final exam
a) Pr(M)=0.30
Pr(F)=0.25

$$\Pr(M \cap F) = 0.20$$



We want to find $\Pr(not\ F/M) = \Pr\ (not\ F \cap M)/\ \Pr\ (M)$

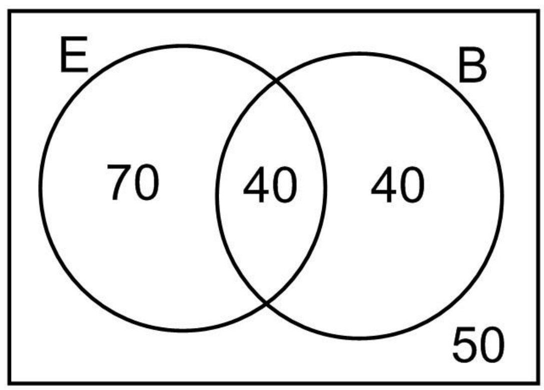
We aren't given any conditional probabilities, so just draw a Venn diagram to solve this
$\Pr(not\ F/M) = 0.10/0.30 = 1/3$

b)$\Pr(not\ M\ and\ not\ F) = 1 - \Pr(M\ or\ F) = 1 - 0.35 = 0.65$
using the Venn diagram above

P43. The total is out of 200 for all fractions.
$\text{n}(E \cup B) = 200 - 50 = 150$

Pr(E)=110/200=11/20
Pr(B)=80/200=2/5

$\text{n}(E\ or\ B) = \text{n}(E) + \text{n}(B) - \text{n}(E\ and\ B)$
150=110 +80 – $n(E \cap B)$
$n(E\ and\ B) = 40$



$\Pr\big((not\ E\ and\ not\ B)\big) = \frac{50}{200} = 1/4$

a) Pr(E/B)=$\frac{\Pr\ (E\ and\ B)}{\Pr\ (B)} = \frac{40/200}{80/200} = 1/2$

b) Pr(E or B but not both)= (70+40)/200=110/200=11/20
NOTE: you don't include the middle of the Venn

P44.

$$\Pr(\text{A/B}) = \frac{\Pr(AandB)}{\Pr(B)} = \frac{0.2}{0.6} = 1/3$$

P45.
$\Pr(E/F)=\frac{\Pr(E \text{ and } F)}{\Pr(F)}$

$2/3=\frac{Pr(E\cap F)}{1/3}$

$\Pr(E \text{ and } F) = 2/9$

The answer is a).

P46.

$\Pr(E/F)=\frac{\Pr(E\text{and } F)}{\Pr(F)}$

$$0.40 = \frac{0.2}{\Pr(F)}$$

Pr(F)=1/2

P47.
If A and B are independent, $\Pr(A \text{ and } B) = \Pr(A) \times \Pr(B)$=0.30x0.20=0.06
Pr(A or B)=Pr(A) + Pr(B) – Pr(A and B)=0.3 + 0.2 - 0.06 = 0.44

P48. They are mutually exclusive, so there is no overlap of circles
Pr(B) = 1 - 0.3 - 0.25 = 0.45

The answer is b).

P49.

		Smoking Status		
		Nonsmoker	*Moderate Smoker*	*Heavy Smoker*
Hypertension Status	*Hypertension*	21	36	30
	No Hypertension	48	26	19

(a) What is the probability that a randomly selected individual is experiencing hypertension?

$$\Pr(\text{hypertension}) = \frac{\#\text{with hypertension}}{\text{total }\#} = \frac{21+36+30}{180} = \frac{87}{180} \approx 0.48$$

(b) Given that a heavy smoker is selected at random from this group, what is the probability that the person is experiencing hypertension?

$$Pr(\text{hypertension}|\text{heavy smoker}) = \frac{Pr(\text{hypertension and heavy smoker})}{Pr(\text{heavy smoker})}$$
$$= \frac{30}{30+19} \approx 0.61$$

(c) Are the events "hypertension" and "heavy smoker" independent? Give supporting calculations.

Since $\Pr(\text{hypertension} \mid \text{heavy smoker}) = \frac{30}{49} \neq \frac{87}{180} = \Pr(\text{hypertension})$, the two events are *not* independent.

P50. **(a)** Are the events *A* and *B* disjoint? Explain.

Yes. They are disjoint because an adult cannot have a college level education and have his highest level of education be secondary.

(b) Are the events *A* and *C* disjoint? Explain.
No. They are not disjoint since females can have a college level education.

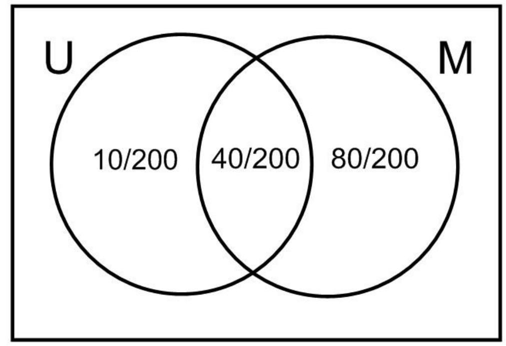
(c) What is the probability that an adult selected at random either has a college level education or is female?
Pr(college or female) = Pr(college) + Pr(female) - Pr(college and female)
$$= \frac{22+17}{200} + \frac{45+50+17}{200} - \frac{17}{200} = \frac{39}{200} + \frac{112}{200} - \frac{17}{200} = \frac{134}{200} = 0.67.$$

(d) What is the probability that an adult selected at random has a college level education given that the adult is a female?
$$\Pr(\text{college} \mid \text{female}) = \frac{\Pr(\text{college and female})}{\Pr(\text{female})} = \frac{\frac{17}{200}}{\frac{112}{200}} = \frac{17}{112} \approx 0.15$$

(e) Are the events *A* and *C* independent?
$$\Pr(A) = \frac{22+17}{200} = \frac{39}{200}; \qquad \Pr(C) = \frac{45+50+17}{200} = \frac{112}{200} = \frac{14}{25};$$
$$Pr(A \text{ and } C) = \frac{17}{200} = 0.085; \qquad \Pr(A)\Pr(C) = \frac{39}{200}\cdot\frac{14}{25} = \frac{273}{2500} = 0.1092;$$
Since $Pr(A \text{ and } C) \neq Pr(A)\,Pr(C)$, the events *A* and *C* are not independent.

P51.
Pr(Female and lower division)=1 - 130/200 = 70/200

Pr(lower division /female)=$\frac{\Pr(both)}{\Pr(female)} = \frac{70/200}{80/200} = \frac{70}{80} = 0.875$

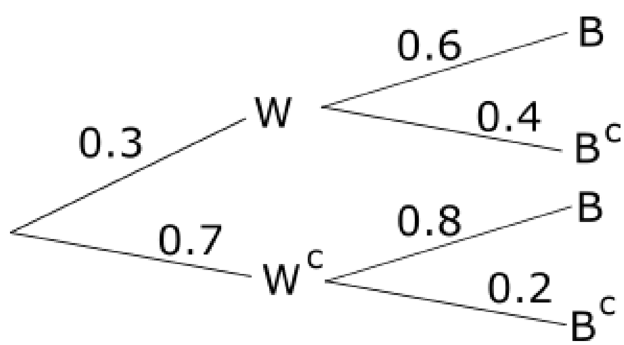


P52.
Pr(fail stop/not signal)= $\frac{\Pr(fail\ stop\ and\ not\ signal)}{\Pr(not\ signal)} = \frac{0.10}{0.15} = \frac{10}{15} = \frac{2}{3}$=0.67

P53.

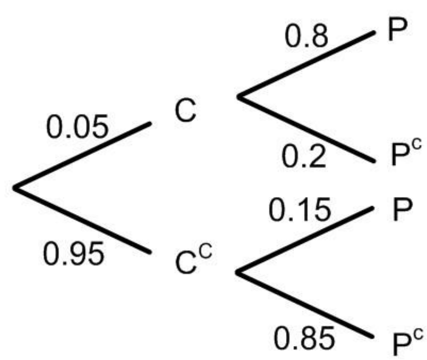
Pr(B)= 0.3(0.6) + (0.70)(0.80)= 0.18 + 0.56 = 0.74

P54.
Pr(W^C/B)=$\frac{\Pr(W^C\ and\ B)}{\Pr(B)} = \frac{0.7(0.8)}{0.74} = \frac{56}{74} = 0.76$



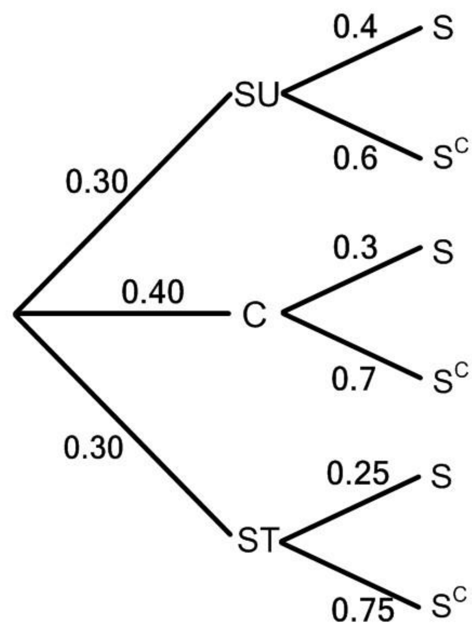
P55. **NOTE:** $C^C = C' = free\ of\ cancer$

Draw a Tree diagram



Pr(C^C/P)=$\frac{\Pr(C^C\ and\ P)}{\Pr(P)} = \frac{0.95(0.15)}{0.95(0.15)+0.05(0.80)}$=0.781

P56. Pr(ST/S)=$\frac{\Pr(ST\ and\ S)}{\Pr(S)} = \frac{0.30(0.25)}{0.3(0.25)+0.4(0.3)+0.3(0.4)}$=0.238



P57. The following two-way table shows the age and sex of all undergraduate university students at a particular university.

Age Group	Female	Male	Total
15-17 years	200	250	450
18-20	3000	3500	6500
21-26	2000	2500	4500
27-34	800	900	1700
35+	500	300	800
Total	6500	7450	13950

Let A= student chosen at random is female
B= student chosen at random is over 26 years old

Find each of the following:

a) Pr(A and B)
=1300/13950=0.093 or 9.3%

b) Pr(A/B) =Pr(female/over 26)= only look at those over 26 and circle number of females
=$\frac{800+500}{1700+800} = \frac{1300}{2500} = 0.52\ or\ 52\%$

c) Pr(B/A)= Pr(over 26/female)= only look at females and circle those who are over 26
$= \frac{800+500}{6500} = \frac{1300}{6500} = 0.2\ or\ 20\%$

d) Pr $(not\ A/B)$= Pr(not female/over 26 years old)= only look at people over 26 years old and circle the men
=$\frac{900+300}{1700+800} = \frac{1200}{2500} = 0.48\ or\ 48\%$

P58. If $\Pr(A\ or\ B) = 0.7\ and\ \Pr(A) = 0.5\ and\ \Pr(not\ B) = 0.6,\ find\ \Pr(A\ and\ B)$.
Pr(B)=1 – Pr(not B)=1 - 0.60= 0.40

$PrA\ or\ B) = \Pr(A) + \Pr(B) - \Pr\ (A\ and\ B)$

0.70 = 0.50 + 0.40 - Pr($A\ and\ B$)

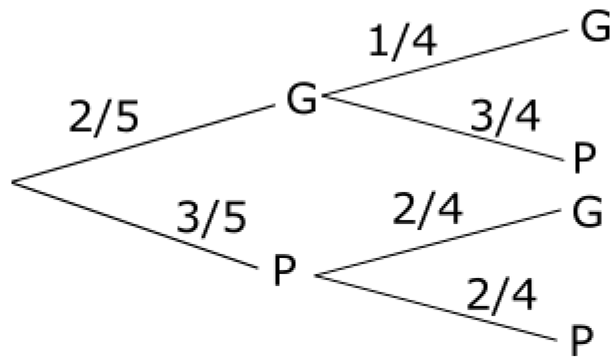
Pr(A $and\ B) = 0.20$

P59.
We want Pr(Cat/Dog)=$\frac{\Pr\ (both)}{\Pr\ (dog)} = \frac{0.18}{0.45} = \frac{18}{45} = \frac{2}{5}$=0.4

P60.
a) What is the probability they are both purple?

Pr(PP)=$\frac{3}{5} \times \frac{2}{4}$=3/10 or 0.3

or $\frac{\binom{3}{2}}{\binom{5}{2}} = \frac{3}{10}$=0.3



b) What is the probability the second pencil is purple if you know the first pencil was green?

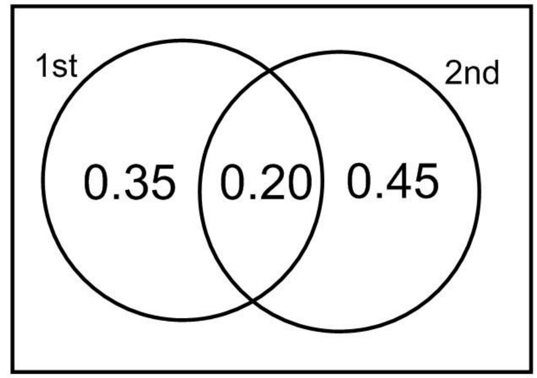
Pr(2nd P/1st G)...draw a tree or use the conditional formula, or both.

OR reduce your sample space...once you take out a green, there will be 1 green and 3 purple left.

So, the prob. the 2nd is purple is 3/4.

c) Pr(different colour)=Pr(GP) + Pr(PG)
$= \frac{2}{5}\left(\frac{3}{4}\right) + \frac{3}{5}\left(\frac{2}{4}\right) = \frac{12}{20} = \frac{3}{5}$

P61. There are no conditional probabilities given, so we use a Venn diagram and not a tree.



55%-20%=35% only winning 1st contract

65%-20%=45% only winning 2nd contract

Prob. of not winning either= outside of the circles in the Venn diagram= 1 - 0.35- 0.20- 0.45=0%

P62. A and B are independent, so Pr(A and B)=Pr(A)xPr(B)

Pr(B/A)=$\frac{\Pr(A \text{ and } B)}{\Pr(A)} = \frac{\Pr(A)\times\Pr(B)}{\Pr(A)} = \Pr(B) = 1/7$

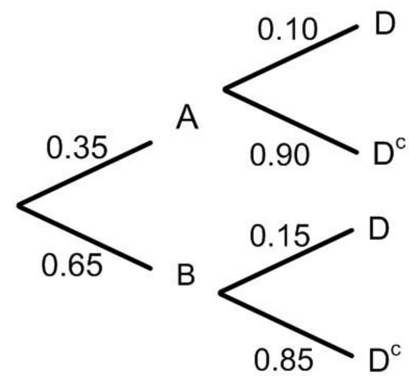
So, Pr(B)=1/7. Recall, if A and B are independent Pr(B/A)=Pr(B) and Pr(A/B)=Pr(A).

P63.

Pr(A or B)=Pr(A) + Pr(B) - Pr(A and B)
0.70=0.4 + 0.5 - Pr(A and B)
Pr(A and B)=0.2

Pr(B/A)=$\frac{\Pr(A \text{ and } B)}{\Pr(A)} = \frac{0.2}{0.4} = 0.5$

P64.



$$\Pr(A/D) = \frac{\Pr(A \text{ and } D)}{\Pr(D)} = \frac{0.35(0.10)}{0.35(0.10) + 0.65(0.15)} = 0.264$$

P65.

Pr(S)=0.20

Pr(D)=0.50
Pr(D/S)=0.85

Pr (D/S)=$\frac{\Pr(D \text{ and } S)}{\Pr(S)}$

Pr(D and S)=(0.85)(0.20)=0.17

P66. The following data represent data for the number of men and women who smoke from a survey.

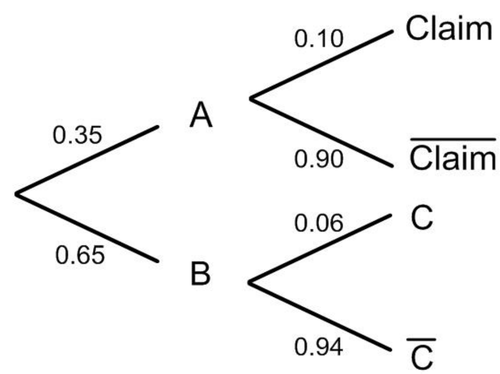
Sex	Smoker	Non-smoker	Total
Male	30	50	80
Female	20	60	80
Total	50	110	160

a) Find the probability a randomly selected female is a smoker.

Pr(smoker/female)= $\frac{\Pr\,(female\ smoker)}{\Pr\,(female)} = \frac{20/160}{80/160} = \frac{20}{80} = \frac{1}{4}\, or\ 0.25$

b) Given that a randomly selected person is a smoker, what is the probability they are male?

Pr(male/smoker)= $\frac{\Pr\,(male\ smoker)}{\Pr\,(smoker)} = \frac{30/160}{50/160} = \frac{30}{50} = 0.60$



P67.

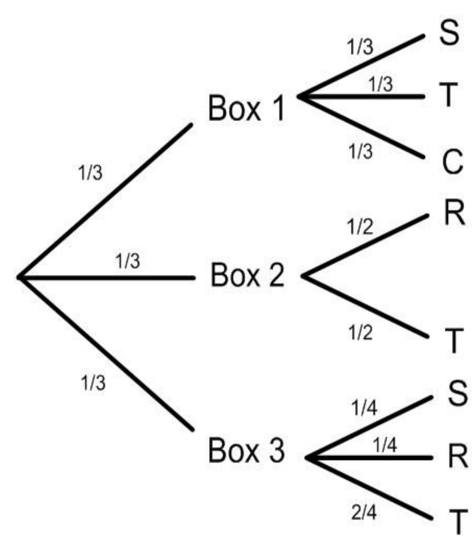
$$\Pr(A/C) = \frac{\Pr\,(A\ and\ C)}{\Pr\,(C)} = \frac{0.35(0.10)}{0.35(0.10)+0.65(0.06)} = 0.473\ \ or\ \ 47.3\%$$

P68.

a) $\Pr(T) = Tr(Box\ 1\ and\ T) + \Pr(Box\ 2\ and\ T) + \Pr(Box\ 3\ and\ T)$

$= \frac{1}{3}\times\frac{1}{3} + \frac{1}{3}\times\frac{1}{2} + \frac{1}{3}\times\frac{2}{4}$

$= \frac{1}{9} + \frac{1}{6} + \frac{1}{6} = \frac{4}{36} + \frac{6}{36} + \frac{6}{36} = \frac{16}{36} = \frac{8}{18} = \frac{4}{9}$=0.44



b) $\Pr(2nd|T) = \frac{\Pr(2nd\ and\ T)}{\Pr(T)} = \frac{1/3 \times 1/2}{4/9} = \frac{1/6}{4/9} = \frac{1}{6}\left(\frac{9}{4}\right) = \frac{9}{24} =$

$\frac{3}{8}$=0.375

P69. Fill in the chart

Forecast snow	76	136	212
Forecast no snow	24	264	288
Total	100	400	500

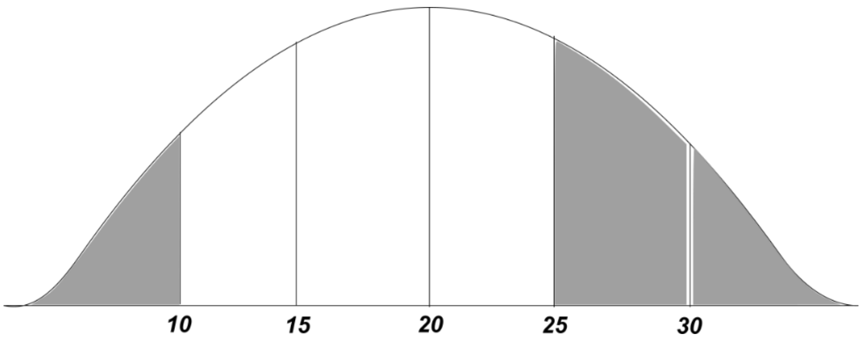
Pr(forecast no snow/ snow)=24/100 since the bottom is only the numbers when it snowed ie. 76+24=100 and the top is forecasted no snow which is 24

You can also use the conditional formula and get

$$\frac{\Pr(forecast\ no\ snow\ and\ it\ snows)}{\Pr(snows)} = \frac{\frac{24}{500}}{\frac{100}{500}} = 0.24$$

P70. from the mean of 20 to 25 is one standard deviation, so on the right side is 68%/2=34% from 20 to 25, so above 25 would be 50%-34%=16%

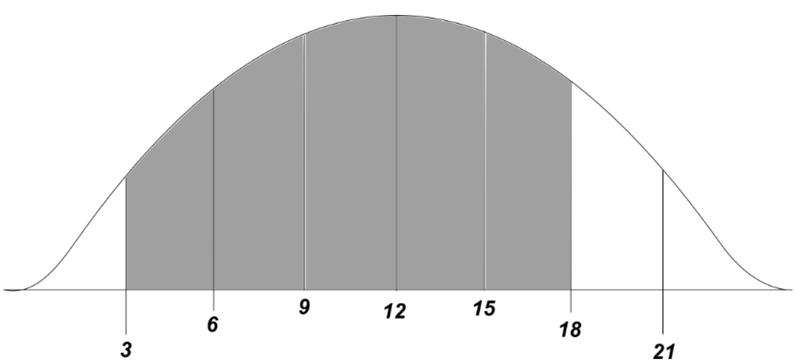
On the left side we want below 10...from 10 to 20 would be 95%/2=47.5, so below 10 would be 50-47.5%=2.5%
The total would be 18.5% for both sides



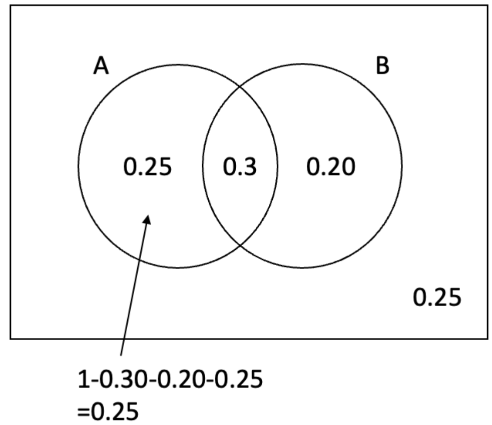
P71. from the mean of 12 up to 18 is two standard deviations, so on the right we have 95%/2=47.5% shaded

on the left we want from 3 to 12, which is 3 standard deviations which would be 99.7/2%=49.85%

The total shading on both sides would be 49.85%+47.5=97.35%



P72.
$\Pr(A \cap B) = 0.3$
$\Pr(A^c \cap B^C) = 0.25$



$\Pr(A \cup B^C)$
$= \Pr(A) + \Pr(B^c) - \Pr(A \cap B^c)$
$= 0.55 + 0.50 - 0.25$
$= 0.80$

P73. $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$
$0.9 = 0.7 + 0.3 - \Pr(A \cap B)$
$\Pr(A \cap B) = 0.1 \neq 0$ ∴ <u>not</u> mutually exclusive

Check independent
Is $\Pr(A \cap B = \Pr(A) \times \Pr(B)$?
$0.1 \neq 0.7 \times 0.3$
$0.1 \neq 0.21$ ∴ not independent either

∴ D) is correct

P74. $\Pr(E^c) = 1 - 0.45 = 0.55$

$\Pr(F) = 1 - 0.20 = 0.80$

$\Pr(E^c \cap F^C) = \Pr(E^C) \times \Pr(F^C)$ since E, F are independent
$= 0.55 \times 0.20 = 0.11$

$\Pr(E^c \cup F^C) = \Pr(E) + \Pr(F^C) - \Pr(E \cap F^C)$
$\Pr(E) + \Pr(F^C) - \Pr(E)\Pr(F^C)$
= 0.45 +0.20 – 0.45(0.20)
=0.65-0.09
=0.56

P75. Since Alexandra is 3 times as likely as Caitlin, if we let Caitlin's probability be x to win, then Alexandra's probability is 3x. Sophie is 4 times as likely to win as Caitlin, since Caitlin's probability o win is ¼ that of Sophie. This way, we avoid fractions!

Event	Probability
Alexandra	3x
Caitlin	x
Sophie	4x
TOTAL	1

The total probability must add up to 1, so we get:

3x+x+4x=1

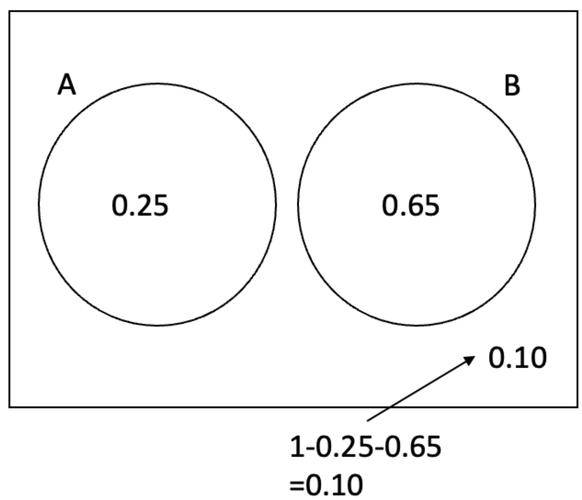
8x=1

x=1/8 and the probability that Alexandra wins is 3/8.

P76.

a)

$= \Pr(A) + \Pr(B) - \Pr(A \cap B)$
$= 0.25 + 0.65 - 0$
$= 0.90$

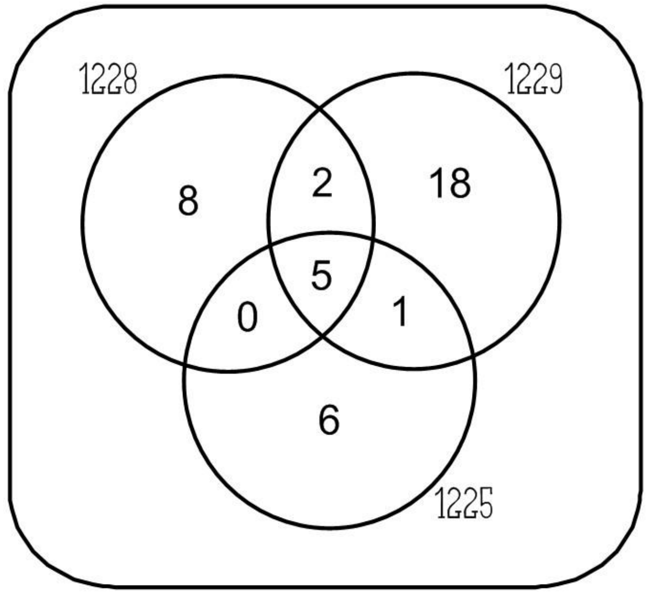


b)

$= \Pr(A) = 0.25$
Since all of A is outside of B

$c)\ = 0.10$
d) = Pr(B) = 0.65 since all of B is out:

P77. a) Number taking exactly one math=8+18+6 = 34



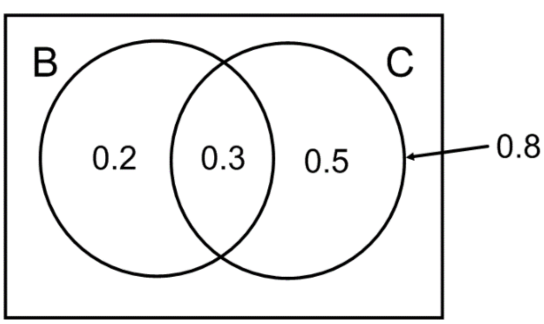
b) Pr(math 1228 and math 1229)= (5+2)/ 80 = 7/80

c) Pr(none of these three) = 1 – (8+2+5+0+18+1+6)/ 80
= 1 – 40/80 = ½

d) Pr (taking only math 1225) = 6/80 = 3/40

P78.You can draw a Venn diagram.

$$\therefore \Pr(exactly\ 1) = 0.5 + 0.2$$
$$= 0.7$$



P79.

Pr(sum 12 given same #)=Pr{(6,6,6)}/Pr$\{(1,1,1)(2,2,2)\big)(3,3,3)(4,4,4)(5,5,5,)(6,6,6)\}$

$=\frac{\frac{1}{216}}{\frac{6}{216}} = \frac{1}{6}$

or reduce sample space S={(1,1,1)(2,2,2)(3,3,3**,)(4,4,4**)(5,5,5)(6,6,6)} since they have to have the same number on all three dice...circle ones with a sum of 12=1/6

P80.

$$\Pr(2nd\ girl\ given\ at\ least\ 1\ boy) = \frac{\Pr(BG)}{1 - \Pr\text{ (no boys)}} = \frac{1/4}{1 - 1/4} = \frac{1}{3}$$

or reduce the sample space S={at least 1 B}={ GB, ***BG,*** BB***}...*** three outcomes and you want the prob. the 2nd is a girl, so, the prob. is 1/3

P81.

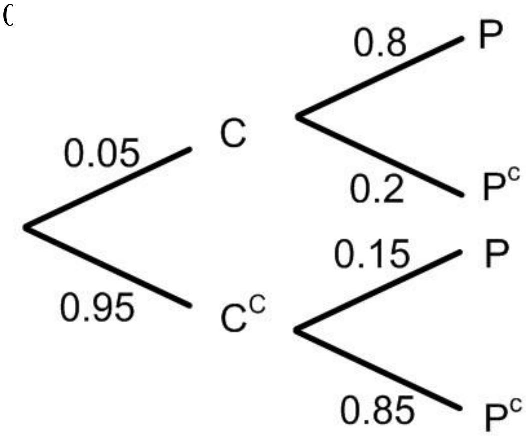
Pr(2 boys given one son)=$\frac{(\Pr 2\ boys\ and\ \ has\ one\ son)}{\Pr\ (has\ a\ son)} = \frac{1/4}{3/4} = \frac{1}{3}$

or reduce the sample space S={GB, BG, ***BB***} since you know they have a son, so it can't be GG...then circle the outcomes with two boys...i.e 1/3

P82. Using Baye's Theorem:

$$\Pr\ (B/A) = \frac{\Pr(A/B)Pr(B)}{\Pr\ (A)} = \frac{\frac{1}{3}(\frac{1}{2})}{\frac{1}{4}} = \frac{1}{6}\left(\frac{4}{1}\right) = \frac{4}{6} = \frac{2}{3}$$

P83. Draw a Tree diagram



$Pr(C^C/P) = \frac{Pr(C^C \text{ and } P)}{Pr(P)} = \frac{0.95(0.15)}{0.95(0.15)+0.05(0.80)} = 0.781$

b) Sensitivity=0.80 (test positive/have disease)
c) Specificity=0.85 (test negative/don't have disease)

Q. Sampling Distributions

Example 1.

A. Sampling distribution

Explanation:
The sampling distribution refers to the distribution of a statistic (in this case, the sample mean) across all possible samples of the same size (100 Canadians). This distribution shows how the statistic (mean) would vary if multiple samples were taken from the population.

Example 2. A study is conducted to investigate the proportion of students who pass an entrance exam at a university. Out of 500 students who take the exam, 375 pass. The study's goal is to compare the proportion of students who pass the exam at this university to the national average, which is 0.80.

Which of the following statements about this information is/are correct? Select all that apply.

A. The value 0.80 is a parameter.
B. The sample size in the study is 500.
C. The value 0.75 (i.e., 375/500) should be considered a statistic.
D. The study used a simple random sampling procedure.

Solution:

A. The value 0.80 is a parameter.
B. The sample size in the study is 500.
C. The value 0.75 (i.e., 375/500) should be considered a statistic.

Explanation:

- **A** is correct because the national average of 0.80 is a fixed value representing the population parameter.
- **B** is correct as the sample consists of 500 students who took the exam.
- **C** is correct since 0.75 is a sample statistic representing the proportion of students who passed the exam in this particular sample.
- **D** is not necessarily correct unless further details about the sampling procedure are provided.

Example 3.

C. A histogram of the average test scores from simple random samples of 30 students each, from a group of 500 students in a Toronto school district.
D. A histogram of the medians of monthly sales figures for simple random samples of 12 stores out of 200 stores in a retail chain.

Explanation:

- **C** and **D** both refer to sampling distributions, where you are looking at a statistic (average test scores or medians) across multiple random samples. Sampling distributions describe the variability of a statistic (e.g., mean or median) from sample to sample.
- **A** is false since it is a distribution of weights (a variable) for a population (all 3rd year university students in engineering in Toronto
- **B** is false because it is a distribution of daily high temperatures for a population, and again NOT from repeated samples, and therefore they do not represent sampling distributions.
- **Note: If they gave an answer that was only from one sample, that would also NOT be a sampling distribution**

Example 4.

The mean of the sample will get closer and closer to the mean of the population, $\mu_{\bar{x}} = \mu = 450$
The standard deviation is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{99}{\sqrt{200}} = 7$

Example 5.

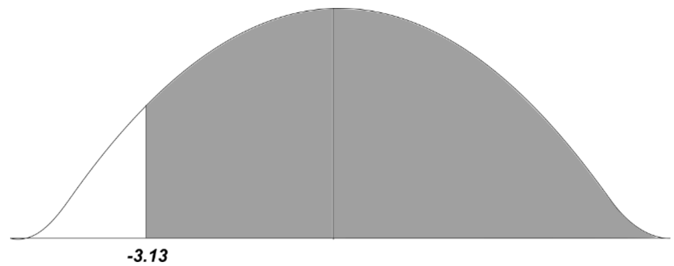
$total\ \ 82\ kg$

$$\bar{x} = \frac{82}{20} = 4.1$$

$$Z = \frac{x-\mu}{\sigma/\sqrt{n}}$$

$$\Pr(\bar{x} < 4.1) = \Pr\left(Z < \frac{4.1-4.2}{1.05/\sqrt{20}}\right)$$
$$= \Pr(Z < -0.43)$$
$$= 0.3336$$

Example 6.



$\mu = 10.6 \qquad \sigma = 0.8 \qquad n = 100$

$$Z = \frac{x - \mu}{\sigma/\sqrt{n}}$$

$$\Pr(\bar{x} > 10.35) = \Pr\left(z > \frac{10.35 - 10.6}{\frac{0.8}{\sqrt{100}}}\right)$$
$$= \Pr(z > -3.13)$$
$$= 1 - 0.0009 = 0.9991$$

Example 7. The answer is D).

Example 8. The answer is B).

Practice Exam Questions on Sampling Distributions

Q1. (a) Let X be the diameter of a ping pong ball. Then $X \sim \mathrm{N}(\mu, \sigma^2)$ with $\mu = 33.0$, $\sigma = 1.0$

$$\Pr(32.5 < X < 33.0) = \Pr\left(\frac{32.5-33.0}{1.0} < Z < \frac{33.0-33.0}{1.0}\right) = \Pr(-0.5 < Z < 0.0)$$

$$= \Pr(Z < 0.0) - \Pr(Z < -0.5) = 0.5000 - 0.3085 = 0.1915.$$

(b) $\Pr(33.3 < X < 33.8) = \Pr\left(\frac{33.3-33.0}{1.0} < Z < \frac{33.8-33.0}{1.0}\right) = \Pr(0.3 < Z < 0.8)$

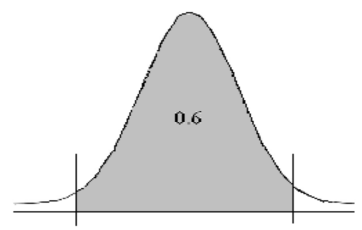
$$= \Pr(Z < 0.8) - \Pr(Z < 0.3) = 0.7881 - 0.6179 = 0.1702.$$

(c) $\Pr(-z < Z < z) = 0.6 \Rightarrow 2\Pr(0 < Z < z) = 0.6$

$$\Rightarrow \Pr(0 < Z < z) = 0.3$$

$$\Rightarrow \Pr(Z < z) = 0.8.$$

$$\Rightarrow z = 0.84.$$



Look up the area below the line to the right (which would be 0.80) and then look up area below the line on the left which would be 0.20 .there is 0.60 in the middle, so 0.20 on each side...total area below the line on the right is 0.6+0.2=0.80...just like it says look up 0.8 in the body and look up the line on the left 0.20 area in the body and find the z scores and they are -0.84 and +0.84 and plug them into the formula with the mean and standard deviation to find the x values

The two z-scores are $z = \pm 0.84$, so since $x = \mu + z\sigma$, the two diameters are

$$\mu - 0.84\sigma = 33.0 - 0.84 \cdot (1.0) = 32.16$$

and $$\mu + 0.84\sigma = 33.0 + 0.84 \cdot (1.0) = 33.84.$$

That is $\Pr(32.16 < X < 33.84) = 0.6$, so 60% of the ping pong balls will have diameters between $32.16\,\text{mm}$ and $33.84\,\text{mm}$.

Q2.(a) Let X be the number of minutes using e-mail. Then $X \sim N(\mu, \sigma)$ with $\mu = 8$, $\sigma = 2$

The sample size is $n = 25$, so $\mu_{\bar{X}} = \mu = 8$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{25}} = 0.40$.

The z-score is given by $Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} < \frac{\bar{X} - 8}{0.40}$. Therefore

$$\Pr(7.8 < \bar{X} < 8.2) = \Pr\left(\frac{7.8 - 8}{0.40} < Z < \frac{8.2 - 8}{0.40}\right) = \Pr(-0.5 < Z < 0.5)$$

$$= \Pr(Z < 0.5) - \Pr(Z < -0.5) = 0.6914 - 0.3085 = 0.3829.$$

(b) The sample size is now $n = 100$, so $\mu_{\bar{X}} = \mu = 8$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{100}} = 0.20$.

The z-score is now given by $Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} < \frac{\bar{X} - 8}{0.20}$.

Therefore,

$$\Pr(7.8 < \bar{X} < 8.2) = \Pr\left(\frac{7.8 - 8}{0.20} < Z < \frac{8.2 - 8}{0.20}\right) = \Pr(-1.0 < Z < 1.0)$$

$$= \Pr(Z < 1.0) - \Pr(Z < -1.0) = 0.8413 - 0.1587 = 0.6826.$$

Q3. The Central Limit Theorem says that the sampling distribution of sample means approaches to a normal distribution $N(\mu, \frac{\sigma}{\sqrt{n}})$ when the sample size gets large.

Q4. **(a)**

Let X be the tuition of an undergraduate student. Then $\mu = \$4172$ and $\sigma = 525$.

$$\Pr(X < 4000) = \Pr\left(Z = \frac{X-\mu}{\sigma} < \frac{4000-4172}{525}\right) = \Pr(Z < -0.33) = 0.3707.$$

(b) $n = 36$, $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 525/\sqrt{36} = 87.5$.

$$\Pr(\bar{X} < 4000) = \Pr\left(Z = \frac{\bar{X}-\mu_{\bar{X}}}{\sigma_{\bar{X}}} < \frac{4000-4172}{87.5}\right) = \Pr(Z < -1.97) = 0.0244.$$

(c) The reason that the probability for part (b) is much lower than that in part (a) is because the sampling distribution of mean in part (b) has much smaller spread with a lot more values distributed near the centre than the population distribution in part (a). While few sample mean values, $\bar{X}$, are lower than 4000, there are many individual values, X, lower than 4000.

Q5. $\mu = 1.5$ $\sigma = 0.5$ $n = 100$

$$Z = \frac{x-\mu}{\sigma/\sqrt{n}}$$

$$\Pr(\bar{x} > 1.0) = \Pr\left(z > \frac{1.0-1.5}{\frac{0.5}{\sqrt{100}}}\right)$$

$= \Pr(z > -10) = 1$. NOTE: The area below -3.49 is almost 0, so the area below -10 is 0 and the area above it would be 1.

The answer is A

Q6. Assume that men's weighs are normally distributed…

$\mu = 172$ and $\sigma = 29, n = 25$

$$Z = \frac{x-\mu}{\sigma/\sqrt{n}}$$

$$\Pr(155 < \bar{x} < 180) = \Pr\left(\frac{155-172}{29/\sqrt{25}} < Z < \frac{180-172}{29/\sqrt{25}}\right) = \Pr(-2.93 < Z < 1.38)$$

Use table 1 and look up the area…Pr(Z<1.38) – Pr(Z<-2.93)=0.9162-0.0017=0.9145

Q7. **(a)**

Let X be the credit card balance. Then $X \sim \mathrm{N}(\mu, \sigma)$ with $\mu = 2780$, $\sigma = 900$. So

$$\Pr(X < 2500) = \Pr\left(Z = \frac{X-\mu}{\sigma} < \frac{2500-2780}{900}\right) = \Pr(Z < -0.31) = 0.3783.$$

(b) Now we are looking at the distribution for the sample mean $\overline{X}$ in a sample of size $n = 25$. Then $\overline{X} \sim \mathrm{N}(\mu_{\overline{X}}, \sigma_{\overline{X}})$ where the mean and standard deviation for $\overline{X}$ are given by $\mu_{\overline{X}} = \mu = 2780$ and $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{900}{\sqrt{25}} = 180$. Therefore

$$\Pr(\overline{X} < 2500) = \Pr\left(Z = \frac{\overline{X}-\mu_{\overline{X}}}{\sigma_{\overline{X}}} < \frac{2500-2780}{180}\right) = \Pr(Z < -1.56) = 0.0594$$

Q8.

Now we are looking at the distribution for the sample mean $\overline{X}$ in a sample of size $n = 10$. Then $\overline{X} \sim \mathrm{N}(\mu_{\overline{X}}, \sigma_{\overline{X}})$ where the mean and standard deviation for $\overline{X}$ are given by $\mu_{\overline{X}} = \mu = 625$ and $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{150}{\sqrt{10}} = 47.43$. Also, the total is given, so we must divide \$7000 by 10 to get a mean of 700.

Therefore,

$$\Pr(\overline{X} > 700) = \Pr\left(Z = \frac{\overline{X}-\mu_{\overline{X}}}{\sigma_{\overline{X}}} > \frac{700-625}{47.73}\right) = \Pr(Z > 1.58) = 1 - 0.9429 = 0.0571.$$

Q9. $\bar{x} = 112.8 \;, n = 9$

$H_0: \mu = 100$

$H_a: \mu > 100 \quad (1 - sided) \qquad \sigma = 15$

$z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{112.8-100}{\frac{15}{\sqrt{9}}} = 2.56$

$p - value = 1 - 0.9948 = 0.0052 < 1\%$

$\therefore p162 \; reject \; H_0 \quad \therefore there \; is \; evidence \; of \; principals \; claim$

R. Confidence Interval for a Mean

Example 1.

(a) Set up a 99% confidence interval for the true population mean amount of paint contained in 5-litre cans.

$\sigma = 0.1$, $n = 50$, $\bar{x} = 4.975$.

$\alpha = 0.01$, $z^* = 2.576$.

The 99% confidence interval for μ is

$$\mu = \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = \bar{x} \pm 2.576 \frac{\sigma}{\sqrt{n}} = 4.975 \pm 2.576 \cdot \frac{0.1}{\sqrt{50}} = 4.975 \pm 0.0364$$

$$= (4.939, 5.011).$$

(b) No, the manager cannot complain to the manufacturer because he can be 99% certain that the true mean lies between $4.939\,\text{L}$ and $5.011\,\text{L}$.

Example 2.

$$\bar{x} = \frac{18.52 + 21.48}{2} = 20$$

width $= 21.48 - 18.52 = 2.96$, so m=width/2 = 2.96/2 = 1.48

$m = Z^* \left(\frac{\sigma}{\sqrt{n}}\right)$

$1.48 = 1.96(\frac{\sigma}{\sqrt{50}})$

$1.48 = 0.277186\sigma$

$\sigma = 5.34$

Example 3.

$$Z^* = 1.645 \; for\ a\ 90\% \; confidence\ interval$$

New error= $1.48 \div 1.96 \times 1.645$

$= 1.24$

$new\ M = (\bar{x} - M, \bar{x} + M)$

$= (20 - 1.24, 20 + 1.24)$

$= (18.76,\ 21.24)$

Example 4.

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{9}} = 1.67$$

The answer is C).

Example 5.

$$n = 15 \quad \bar{x} = 47 \quad \sigma = 5 \quad 90\%\ CI \quad Z^* = 1.645$$

$$\mu = \bar{x} \pm z^* \left(\frac{\sigma}{\sqrt{n}}\right) = 47 \pm 1.645\left(\frac{5}{\sqrt{15}}\right)$$

$$= 47 \pm 2.12 = (44.88, 49.12)$$

Example 6.
n=25
$\bar{x} = 450$

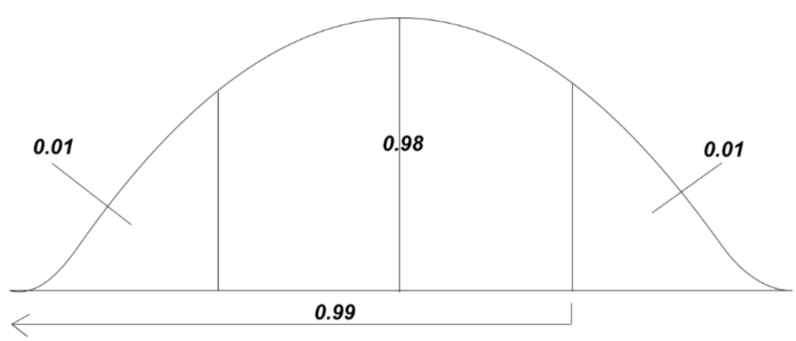
90% CI is $\mu = \bar{x} \pm z^* \left(\frac{\sigma}{\sqrt{n}}\right)$

If we find a 95% confidence interval, the margin of error, $z^* \left(\frac{\sigma}{\sqrt{n}}\right)$ will be larger since the value of z^* will be greater

If we use a smaller sample size, n, we will divide by a smaller number and therefore, the margin of error will increase as well.

The answer is E.

Example 7.
a) $m = z^* \left(\frac{\sigma}{\sqrt{n}}\right) = 2.326 \left(\frac{4}{\sqrt{100}}\right) = 0.9304$

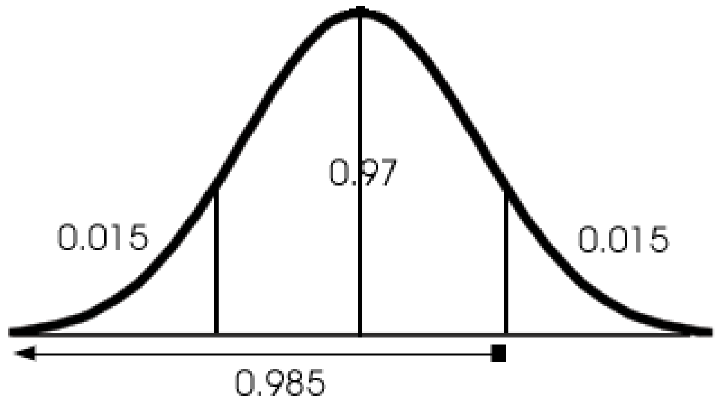


Look up Area 0.99 in the body to get $Z^* = 2.33$

b) m= width/2

width= margin of error (2) = 0.9304 (2) = 1.8608

Example 8. 97% CI



Look up
0.985 $in\ body\ \ Z_{crit} = z^* = \pm 2.17$

Example 9. If the width is 41.216, the margin of error is 41.216/2=20.608

$m = Z^*\left(\frac{\sigma}{\sqrt{n}}\right)$

$20.608 = Z^*(8)$

$Z^* = 2.576, so\ it\ is\ a\ 99\%\ confidence\ interval$

Bootstrap Confidence Intervals

Example 10.

100%	Maximum	2.74
99.5%		2.42
97.5		2.05
95		1.83
90		1.75
50	Median	1.31
25		1.13
10		0.99
5		0.91
2.5		0.86
0	Minimum	0.65

The 95% confidence interval would be:
(0.86, 2.05)
We are 95% confident the true population mean is in this interval.

The 90% confidence interval would be:
(0.91, 1.83)

We are 90% confident the true population mean is in this interval.

Example 11.

The confidence interval is (27.16, 30.01)

This means we are 95% confident the true population mean lies in this interval.

So, if the hypothesis is the mean is equal to 29, what would your conclusion be?

Since 29 is in our 95% confidence interval, we would not reject the hypothesis. (statistically insignificant)

What is the hypothesis is the mean is equal to 32, what would your conclusion be? (statistically significant)

Since 32 is not in our 95% confidence interval, we would reject the hypothesis.

Practice Exam Questions on Confidence Intervals

R1. (a)

$$\bar{x} = \frac{190.5+189.0+195.5+187.0}{4} = \frac{762}{4} = 190.5.$$

$\sigma = 3.14$, $n = 4$.

$\alpha = 0.10$, $z^* = 1.645$.

The 90% confidence interval for μ is

$$\mu = \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = \bar{x} \pm 1.645 \frac{\sigma}{\sqrt{n}} = 190.5 \pm 1.645 \cdot \frac{3.14}{\sqrt{4}} = 190.5 \pm 2.58$$

$$= (187.9, 193.1).$$

(b) The 90% confidence interval for μ would now be

$$\mu = \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = \bar{x} \pm 1.645 \frac{\sigma}{\sqrt{n}} = 190.5 \pm 1.645 \cdot \frac{3.14}{\sqrt{1}} = 190.5 \pm 5.17$$

$$= (185.3, 195.7).$$

(c) α has been changed from $\alpha = 0.10$ to $\alpha = 0.01$, so now $z^* = 2.576$.

The 99% confidence interval for μ is therefore

$$\mu = \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = \bar{x} \pm 2.576 \frac{\sigma}{\sqrt{n}} = 190.5 \pm 2.576 \cdot \frac{3.14}{\sqrt{4}} = 190.5 \pm 4.04$$

$$= (186.5, 194.5).$$

R2 n=14., $\bar{x} = 65.12$, $\sigma = 24.60$

$\alpha = 0.05$,

The 95% confidence interval for μ is

$$\mu = \bar{x} \pm Z^* \left(\frac{\sigma}{\sqrt{n}}\right) = 65.12 \pm 1.96 \left(\frac{24.6}{\sqrt{14}}\right) = 65.12 \pm 12.89 = (52.23, 78.01)$$

R3. n=15., $\bar{x} = 29.50$, $\sigma = 12$

$\alpha = 0.05$,

The 95% confidence interval for μ is

$$\mu = \bar{x} \pm Z^* \left(\frac{\sigma}{\sqrt{n}}\right) = 29.50 \pm 1.96 \left(\frac{12}{\sqrt{15}}\right) = 29.5 \pm 6.07 = (23.43, 35.57)$$

R4. $n = 20$, $\bar{x} = 1.67$, $\sigma = 0.32$.

$\alpha = 0.05$, $z^* = 1.96$

The 95% confidence interval for μ is

$$\mu = \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = \bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} = 1.67 \pm 1.96 \cdot \frac{0.32}{\sqrt{20}} = 1.67 \pm 0.14$$

R5.

$n = 10$, $\bar{x} = 261$, $\sigma = 139$.

$\alpha = 0.10$, $z^* = 1.645$.

The 95% confidence interval for μ is

$$\mu = \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = \bar{x} \pm 1.645 \frac{\sigma}{\sqrt{n}} = 261 \pm 1.645 \cdot \frac{139}{\sqrt{10}} = 261 \pm 72.3$$

R6. (a)

$$\bar{x} = 540$$
$$\sigma = 80$$
$$n = 10$$

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = 540 \pm 1.96 \frac{80}{\sqrt{10}}$$
$$= 540 \pm 49.58$$
$$= (490.42,\ 589.58)$$

(b)

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = 540 \pm 2.575 \frac{80}{\sqrt{10}}$$
$$= 540 \pm 65.14$$
$$= (474.86,\ 605.14)$$

This interval is wider since in order to be more confident that the interval contains the true population mean, we need a larger range at values.

R7. $\sigma = 16, n = 15, \bar{x} = 105, 95\%\ CI$

$$\mu = \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$
$=105 \pm 1.96 \frac{(16)}{\sqrt{15}}$
$=105 \pm 8.097$
=(96.903, 113.097)

R8. (a) $\sigma = 100$, $n = 64$, $\bar{x} = 350$.

$\alpha = 0.05$, $z^* = 1.96$.

The 95% confidence interval for μ is

$$\mu = \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = \bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} = 350 \pm 1.96 \cdot \frac{100}{\sqrt{64}} = 350 \pm 24.5 = (325.5, 374.5)$$

(b) No, since the 95% confidence interval does not contain the claimed mean of 400 hours. We are 95% certain that the true mean lifetime is less than 400 hours.

R9. SE=$\frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{35}} = 0.68$

R10. (56, 68) is the confidence interval, so the width of the interval is 68 -56 = 12 and the margin of error is 1/2 width=1/2 (12) =6

R11. $\bar{x} = 36\ and\ m = 1.3$

Use the formula
New m= old m $\div\ old\ Z^* \times new\ Z^* = 1.3 \div 2.576 \times 1.645 = 0.83$

R12. a) look up the area 0.97 (0.94 total on both sides around 0 and so there is 0.0.3 on each side) on the Z-table...you get Z=1.88

b) look up the area 0.94 on the Z-table...you get Z=1.555

R13. $Z^* = 2.33$

Look up area below $2.33\ and\ get$ 0.9901
∴ area to the right is 0.01 ∴ both sides are 0.01 × 2
∴ $98\%\ CI$

R14. $95\%\ CI\ \ (86.45, 89.49)$

$$m = \frac{1}{2} width = \frac{1}{2}(89.49 - 86.49)$$
$$= \frac{1}{2}(3) = 1.5$$
$$\bar{x} = \frac{86.49+89.49}{2} = 87.99$$
$$\therefore 1.5 \div 1.96 \times 2.576 = 1.97 \quad new\ m$$

$$\therefore (87.99 - 1.97, 87.99 + 1.97)$$

$$= (86.02, 89.96)$$

R15. $90\%\ [800, 900]$ as $z_{\alpha/2} \uparrow\ E \uparrow$
$96\%\ \rightarrow E \uparrow$
The answer is D).

R16. 90% CI $m = 10$ $n = 600$

m= $Z^*(\frac{\sigma}{\sqrt{n}})$

$10 = 1.645(\frac{\sigma}{\sqrt{600}})$

$\sigma = 148.9$

R17. 98% $Z^* = 2.326$

$(47.65,\ 56.35)$

$m = \frac{56.35-47.65}{2} = 4.35$

To find the new Z^*, look up area 0.96 in the body and get 1.75

use the formula

New m= old m $\div\ old\ Z^* \times new\ Z^* =$

$new\ m = 4.35 \div 2.326 \times 1.75$

$= 3.27$

$\bar{x} = \frac{47.65+56.35}{2} = 52$

$new\ CI = (\bar{x} - m,\ \bar{x} + m)$

$\therefore new\ CI = (52 - 3.27,\ 52 + 3.27)$

$(48.73, 55.27)$

R18. I and III are true. So, the answer is B).

S. Finding the Sample Size and Margin of Error

Example 1.

99% $\therefore Z^* = 2.576$

$width = 6 \quad \therefore m = 3$

$\sigma^2 = 35 \quad \therefore \sigma = \sqrt{35} = 5.916$

$n = \left(\frac{Z^*\sigma}{m}\right)^2 = \left(\frac{2.576(5.916)}{3}\right)^2 = 25.8$

$\therefore n = 26 \quad (round\ up)$

The answer is E).

Example 2.

$\sqrt{4} = 2\ times$

twice as wide, which means 4 times smaller in terms of sample size

The answer is A).

As n increases by 4 times, m decreases by $\sqrt{4} = 2\ times$

As n decreases by 4 times, m increases $\sqrt{4} = 2\ times$

Example 3.

95% $CI \quad Z^* = 1.96 \quad \sigma = 10 \quad n = 26$ width=?

$m = Z^*\left(\frac{\sigma}{\sqrt{n}}\right)$

$m = 1.96\left(\frac{10}{\sqrt{26}}\right) = 3.84$

width=$m \times 2 = 7.7$

Practice Exam Questions on Sample Size

S1. A) $\sigma = 54, \mathrm{m} = 8$ (within 8 minutes)

$\alpha = 0.10, so\ z^* = 1.645$

n=? $n = \left(\frac{Z^*\sigma}{m}\right)^2 = (\frac{1.645(54)}{8})^2 = 123.3$

Therefore, 124 is the sample size needed.

b) $\alpha = 0.01, so\ z^* = 2.576$

n=? $n = \left(\frac{Z^*\sigma}{m}\right)^2 = (\frac{2.576(54)}{8})^2 = 302.3$

Therefore, 303 is the sample size needed.

S2. **(a)** What sample size is needed?

$\sigma = 20$, m=width/2=5/2 = 2.5

$\alpha = 0.01$, $z^* = 2.576$.

$$n = \left(\frac{Z^*\sigma}{m}\right)^2 = (\frac{2.576(20)}{2.5})^2 = 424.7$$

Therefore, a sample of size 425 would be required.

(b) If 95% confidence is desired, what sample size is necessary?

$\alpha = 0.05$, $z^* = 1.96$.

$$n = \left(\frac{Z^*\sigma}{m}\right)^2 = (\frac{1.96(20)}{2.5})^2 = 245.9$$

Therefore, a sample of size 246 would be required.

S3. $\sigma = 3.14, \text{m} = \frac{1}{2}(4) = 2$

m= 1/2 of the width of the interval

$\alpha = 0.05$, $z^* = 1.96$.

$$n = \left(\frac{Z^*\sigma}{m}\right)^2 = (\frac{1.96(3.14)}{2})^2 = 9.47$$

Therefore, a sample of size 10 would be required, i.e., he would need to weigh himself at least 10 times per month.

S4. $\sigma = 15.8$, $\text{m} = 4$ (within 4 lbs)

$\alpha = 0.10, so\ z^* = 1.645$

n=?

$$n = \left(\frac{Z^*\sigma}{m}\right)^2 = (\frac{1.645(15.8)}{4})^2 = 42.2$$

We'd need a sample size of 43

S5. $\sigma = 80$, $\text{m} = 10$ (within 10 lbs)

$\alpha = 0.10, so\ z^* = 1.645$

$$n = \left(\frac{Z^*\sigma}{m}\right)^2 = (\frac{1.645(80)}{10})^2 = 173.2$$

$\therefore$ We'd need a sample size of 174. .

S6. m= $0.6(within)$ $\sigma = 2.5$ $n = ?$

95%, $Z^* = 1.96$

$$n = \left(\frac{Z^*\sigma}{m}\right)^2 = \left(\frac{1.96(2.5)}{0.6}\right)^2 = 66.7$$

The answer is B).

Data Science Final Exam 1

Multiple choice:1 mark each= 30 marks

1. The answer is C).
2. The answer is C).
3. The answer is A). $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{11-10}{2/\sqrt{20}} = 2.24$

   $\Pr(z > 2.24) = 1 - 0.9875$

   =0.0125
4. The answer is B). $n = 10 \quad \sigma = 20 \quad \bar{x} = 200$

   CI $\mu = \bar{x} \pm z^*(\frac{\sigma}{\sqrt{n}})$

   $= 200 \pm 2.326(\frac{20}{\sqrt{10}})$

   $= 200 \pm 14.7$

   = (185.3, 214.7)
5. The answer is C). Pr(H) = 0.45

   Pr(S) = 0.30

   Pr(D) = 0.20 –not needed

   Pr(S and H)=0.10

   Pr(H/S)=$\frac{\Pr(H \text{ and } S)}{\Pr(S)} = \frac{0.10}{0.30} = 0.33$
6. The answer is A). $\sigma = 5 \quad n = 20$

   S.E. = standard error = $\frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{20}} = 1.12$

7. The answer is D). Pr(A or B) = Pr(A)+Pr(B) – Pr(A and B)

   =0.15+0.55 – 0.15×0.55

   = 0.6175

   Pr(A and B) = Pr(A)× Pr (B) since independent

   = $0.15\times 0.55 = 0.0825$

8.The answer is E). $\hat{y} = 4.5 - 0.3(1) = 4.2 \;\; \therefore C \text{ is true}$

$-0.3(1) = -0.3 \;\; \therefore \downarrow \text{ of } 0.3 \text{ to } GPA \;\; \therefore B \text{ is true}$

9.

Class	High	Medium	Low	Total
1st year	60	40	10	110
2nd year	50	30	20	100
3rd year	20	20	40	80
Total	130	90	70	290

The answer is E). $\frac{60}{90} = 0.67$

10. The answer is D). Pr(A or B) = Pr(A) + Pr(B) – Pr(A and B)

$$= \frac{2}{6} + \frac{2}{6} - 0$$
$$= \frac{4}{6} = \frac{2}{3}$$

Pr(A and B) = 0 since they have no events in common

$\therefore A\ and\ B$ are mutually exclusive. Recall, if A and B are mutually exclusive, they can't be independent.

11. The answer is A). $m = Z^*\left(\frac{\sigma}{\sqrt{n}}\right) = 1.645\left(\frac{7}{\sqrt{50}}\right) = 1.63$

S.E.$= \frac{\sigma}{\sqrt{n}} = \frac{7}{\sqrt{50}} = 0.99$

12.The answer is B). $\bar{x} = \frac{\sum x}{n} = \frac{600\ 000}{15} = 40\ 000$

Pr$(\bar{x} > 40\ 000)$ $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{40\ 000-45\ 000}{6000/\sqrt{15}} = -3.23$

Pr$(z > -3.23) = 1 - 0.0006 = 0.9994$

13. The answer is E). *It doesn't ask about a mean, average or total,
So it is not $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$

$z = \frac{x-\mu}{\sigma} = \frac{7-5.6}{1.8} = 0.78$

$$\Pr(x > 7) = \Pr(z > 0.78)$$
$$= 1 - 0.7823$$
$$= 0.2177$$

$\therefore 0.2177 \times 20 = 4$

14. The answer is E). A. is r^2 so $r = \sqrt{r^2}$

B. is also true

$b = r\,\frac{S_y}{S_x}$ $So, if\ we\ know\ S_y$ & S_x and we have

b = 1.2 from the equation, we can calculate r

15. The answer is D). larger width = larger m

CI $\mu = \bar{x} \pm Z^*(\frac{\sigma}{\sqrt{n}})$

$as\ Z^* \uparrow, m \uparrow$

$as\ \sigma \uparrow, m \uparrow$

$as\ n \downarrow, m \uparrow$

16. The answer is B). width = 55-30 = 25 m = $\frac{width}{2} = 12.5$

$$m = Z^*(\frac{\sigma}{\sqrt{n}})$$

$$12.5 = 1.645(\frac{\sigma}{\sqrt{20}})$$

$$\sigma = 33.98$$

17. The answer is B). $m = \frac{140-60}{2} = 40$

$$n = (\frac{Z^*\sigma}{m})^2 = (\frac{1.96(80)}{40})^2 = 15.4 \ \therefore 16$$

18. The answer is D). Pr(F/E) = $\frac{\Pr(E\ and\ F)}{\Pr(E)} = \frac{0.2}{0.8} = 0.25$

19. The answer is D). r is unitless

20. The answer is A). -5, 1.6, 2.8, $\boxed{3.5}$, 4.2, 5.9, 8.7

Q1 = 1.6

Q3 = 5.9

IQR = Q3 – Q1 = 5.9 – 1.6 = 4.3

Below Q1 – 1.5 IQR

= 1.6-1.5(4.3)

= -4.85, so -5 is an outlier

21. The answer is A). Since x and y are inversely related, as x increases, y decreases, so r must be negative, $r = -\sqrt{0.97} = -\frac{0}{9849}$

$$b = r\frac{S_y}{S_x} = -0.9849\ (\frac{250}{20})$$

$$b = -12.31$$

$$a = \bar{y} - b\bar{x}$$

$$= 300 - (-12.31)(40) = 792.4$$

$$\hat{y} = a + bx$$

$$\hat{y} = 792.4 - 12.31x$$

22. The answer is D). Find Q3

IQR = Q3 – Q1

8000 = Q3 – 3000

Q3 = 11000

Find the Maximum

range = max – min

15 000 = max – 2100

Max = 16 000

Mean>Median, so it is right-skewed.

Maximum, IQR, skew= 16000, 11000, right-skewed

23. The answer is B).

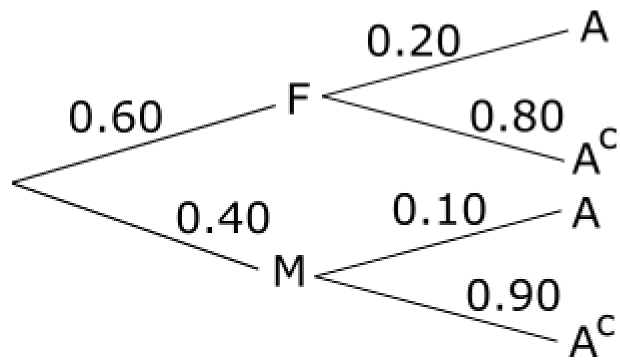
multiply the standardized values together and find it divided by $(n-1)$

The standardized values are $(\frac{x_i-\bar{x}}{s_x}) and\ (\frac{y_i-\bar{y}}{s_y})$

$$n-1=5-1=4$$

$$r=\frac{(1.35)(-0.96)+0.6(1.11)+(0.2)(-0.67)+(-1.04)(-0.57)+(-1.11)(1.09)}{4}=-\frac{1.3811}{4}=-0.345$$

24. The answer is B).



$$\Pr(^{M}/_{A})=\frac{\Pr(M\ and\ A)}{\Pr(A)}=\frac{0.40\times0.10}{0.60(0.20)+0.40(0.10)}$$
$$=\frac{0.04}{0.12+0.04}=\frac{0.04}{0.16}=0.25$$

25. The answer is E). disjoint means Pr(A and B) = 0

∴ C. is true

Pr(A or B) = Pr(A) + Pr(B) – Pr(A and B)

= 0.2 + 0.8 – 0 = 1 ∴ $B.i$s true

26. The answer is B). Pr(at least 1 head) = 1 – Pr(no heads)

= 1 – 0.6× 0.6 × 0.6 × 0.6

= 1 – 0.6^4

= 0.8704

**since trials are independent, we can multiply, and not getting a head each time is the same as getting a tail, i.e. 0.6

27. The answer is A). (62, 84) $m=\frac{84-62}{2}=11$

$$\bar{x}=\frac{62+84}{2}=73$$

28. The answer is A).

Pr(5 or more) = 1 - 0.1 - 0.2 - 0.1 - 0.3 - 0.2 = 0.1

Pr(A) = Pr(3, 4, 5 or more)=0.3+0.2+0.1= 0.6

Pr(B) = Pr(0, 1, 2, 3) = 0.1+0.2+0.1+0.3= 0.7

Pr(A and B) = Pr(3) = 0.30

Pr(B/A) = $\frac{\Pr(A\ and\ B)}{\Pr(A)} = \frac{0.3}{0.6} = 0.5$

29. The answer is D). Pr(A/B) = $\frac{\Pr(A\ and\ B)}{\Pr(B)}$

$0.65 = \frac{\Pr(A\ and\ B)}{0.42}$

Pr(A and B) = 0.273 Pr(B/A) = $\frac{\Pr(A\ and\ B)}{\Pr(A)} = \frac{0.273}{0.62} = 0.4403$

Or use Baye's). Pr(B/A) = $\frac{\Pr(A/B)\Pr(B)}{\Pr(A)} = \frac{0.65(0.42)}{0.62} = 0.4403$

30. The answer is A). $\mu = 50\ is\ not\ in\ (45{,}49)$

$\therefore it\ is\ statistically\ significant$

$\mu = 46\ is\ in\ (45{,}49)$

$\therefore it\ is\ not\ statistically\ significant$

$\therefore i)\ only$

Long answer:

1. $\bar{x} = \frac{\sum x}{n} = \frac{80}{25} = 3.2$

$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ $\Pr(\bar{x} > 3.2) = \Pr\left(z > \frac{3.2-3.5}{1.25/\sqrt{25}}\right)$

$= \Pr(z > -1.2)$

$= 1 - \Pr(z < -1.2)$

$= 1 - 0.1151$

$= 0.8849$

2. a) $\bar{x} = \frac{18+22}{2} = 20$

$m = \frac{22-18}{2} = 2$ 90% $\therefore Z^* = 1.645$

$m = Z^*\left(\frac{\sigma}{\sqrt{n}}\right)$

$2 = 1.645\left(\frac{\sigma}{\sqrt{60}}\right)$

$\sigma = 9.42$

b) $\mu = \bar{x} \pm Z^* \left(\frac{\sigma}{\sqrt{n}}\right) = 20 \pm 1.96 \left(\frac{9.42}{\sqrt{60}}\right) = 20 \pm 2.384$

= (17.616, 22.384)

3. new M = old M $\div\ old\ Z^* \times new\ Z^*$

$= \left(\frac{20-15}{2}\right) \div 1.96 \times 2.576 = 3.29$

$\bar{x} = \frac{15+20}{2} = 17.5$

New CI = $(\bar{x} - M, \bar{x} + M)$

= (17.5 – 3.29, 17.5 + 3.29)

= (14.21, 20.79)

4. a) An experiment; block design (blocked by sex)

b) Explanatory variable is the cholesterol drug
Response variable is the cholesterol levels

c) Lurking variables: diet, exercise as they can affect cholesterol levels and other medications (some depression meds increase cholesterol)

5. a) The number 4 is missing
b) right-skewed

c) 21, 22, 23, 30, **32,** 35, 52, 61, 75

median = 32

Q1 = 22.5

Q3 = $\frac{52+61}{2}$= 56.5

IQR = Q3 – Q1 = 56.5 – 22.5 = 34

d) Q1 – 1.5 IQR

= 22.5 – 1.5 (34) = below -28.5

Q3 + 1.5 IQR

= 56.5 + 1.5 (34) = above 107.5

$\therefore$ *no outliers*

6.

Exposed	Ovarian cancer	No ovarian cancer
Yes	20 a	90 b
No	8 c	100 d

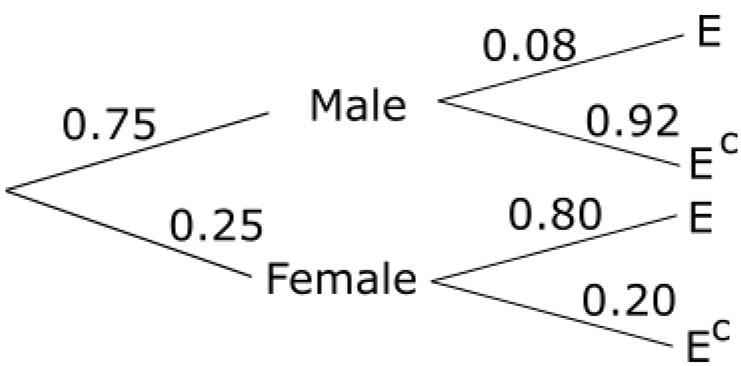
Odds ratio = $\frac{ad}{bc} = \frac{20(100)}{90(8)} = 2.8$

∴ the group of exposed women has 2.8 the odds of getting ovarian cancer than the non-exposed women

Relative risk = $\frac{a(c+d)}{c(a+b)} = \frac{20(8+100)}{8(20+90)} = \frac{20(108)}{8(110)} = 2.5$

∴ the relative risk for the exposed group is 2.5 times that of the non-exposed group

7.a)

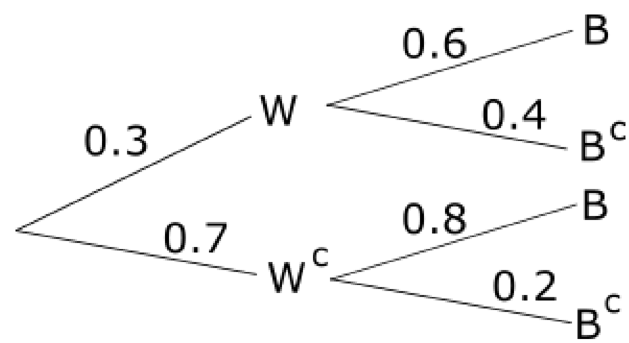


	Everyday E	Not everyday E^c
Male 0.75	0.08	0.92
Female 0.25	0.80	0.20

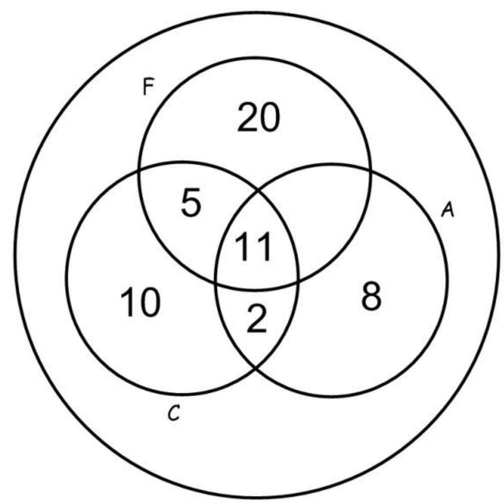
b) Pr(F/E) = $\frac{\Pr(F\ and\ E)}{\Pr(E)} = \frac{0.25(0.80)}{0.25(0.80)+0.75(0.08)} = \frac{0.2}{0.2+0.06} = \frac{0.20}{0.26} = 0.77$

8. a) Pr(B)= 0.3(0.6) + (0.70)(0.80)= 0.18 + 0.56 = 0.74

b) $\Pr(W^C/B) = \frac{\Pr(W^c \text{ and } B)}{\Pr(B)} = \frac{0.7(0.8)}{0.74} = \frac{56}{74} = 0.76$



9.



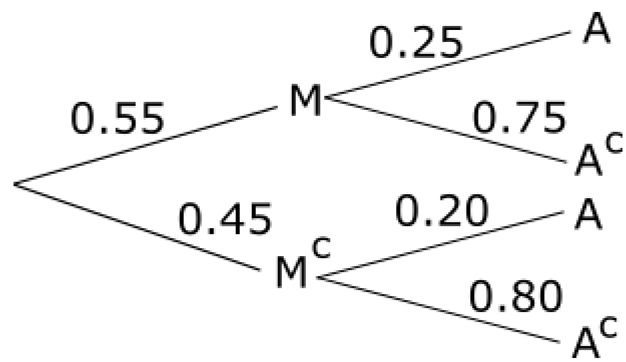
a) From the Venn diagram, 11 students are taking all three courses.

b) Pr(only Finite) = 20/60 = 1/3

c) Pr(Calc and Algebra) = 11 +2 / 60 = 13/ 60

d) Pr(none of these three math classes) = 1 – (20+5+11+2+10+8)/ 60
= 1- 56/60
= 60/60 – 56/ 60
=4/60 or 2/30 or 1/15

10.
$Pr(M/A) = \frac{Pr(M\ and\ A)}{Pr\ (A)} = \frac{0.55(0.25)}{0.55(0.25)+0.45(0.20)} = 0.604\ or\ 60.4\%$



11.
Let X be the tuition of an undergraduate student.
Then $\mu = \$4172$ and $\sigma = 525$.

$$Pr(X < 4000) = Pr\left(Z = \frac{X-\mu}{\sigma} < \frac{4000-4172}{525}\right) = Pr(Z < -0.33) = 0.3707.$$

(b) $n = 36$, $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 525/\sqrt{36} = 87.5$.

$$Pr(\bar{X} < 4000) = Pr\left(Z = \frac{\bar{X}-\mu_{\bar{X}}}{\sigma_{\bar{X}}} < \frac{4000-4172}{87.5}\right) = Pr(Z < -1.97) = 0.0244 .$$

(c) The reason that the probability for part (b) is much lower than that in part (a) is because the sampling distribution of mean in part (b) has much smaller spread with a lot more values distributed near the centre than the population distribution in part (a). While few sample mean values, $\bar{X}$, are lower than 4000, there are many individual values, X, lower than 4000.

Data Science Final Exam 2

1. $z = \frac{x-\mu}{\sigma} = \frac{138-100}{2} = 3.17$

   $\Pr(z < 3.17) = 0.9992$ or 99.92%

   The answer is C).

2. The answer is B).

   Played $\frac{300}{1100} = 0.272$ Don't play $\frac{600}{1200} = 0.5$

3. $\frac{1100}{1200}$ =0.92

   The answer is A).

4. $\hat{y} = 20\ 000 + 900(30) = \$47000$

   Residual= $50 000 - $47,000 = $3000

   The answer is B).

5. 85000 = 20000+900x…it says 85 in 1000's so it is $85,000 we type in:

   65000=900x

   x=72.2

   The answer is E).

6. $z = \frac{x-\mu}{\sigma} = \frac{20-16}{2} = 2$

   $\Pr(z > 2) = 1 - 0.9772$

   $= 0.0228$

   The answer is B).

7. x, -0.9, 0.7, 0.9, 1.2, [1.3], 2.5, 3.6, 4.2, 11.5, 13.8

   Q1 = 0.7 Q3 = 4.2

   IQR = Q3 – Q1 = 4.2 - 0.7 = 3.5

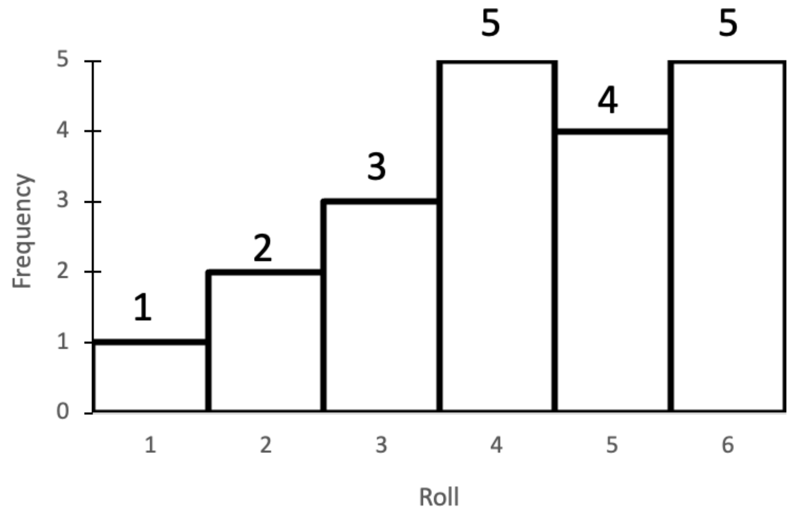
   Q1 – 1.5 IQR = 0.7-1.5(3.5)

   = -4.55 below

   The answer is B).

8. median occurs between 10th and 11th number

$\therefore 1+2+3<10.5$



$1+2+3+5>10.5$, *so the median occurs in the bar with height* 5

$\therefore 4$ *is the median*

The answer is B).

9. The answer is B). New $m = old\ m \div Z^{*(old)} \times Z^{*(new)}$

$= 58 \div 1.645 \times 1.96$

$= 69.1$

10. The answer is D). $\mu = \bar{x} \pm Z^{*}\left(\frac{\sigma}{\sqrt{n}}\right) = 20 \pm 2.576(\frac{6.5}{\sqrt{50}})$

$= 20 \pm 2.4$

$= (17.6, 22.4)$

11. The answer is B). $as\ n \uparrow, width \downarrow$

$10 \times 9 = 90$

$\sqrt{9} = 3$ ∴sample size 10 is 3× as wide as sample size 90

12.

Exposed	Lung cancer	No lung cancer
Yes	25 a	80 b
No	5 c	90 d

The answer is A). odds $= \frac{ad}{bc} = \frac{25(90)}{80(5)} = 5.6$

Relative risk $= \frac{a(c+d)}{c(a+b)} = \frac{25(5+90)}{5(25+80)} = \frac{2375}{525} = 4.5$

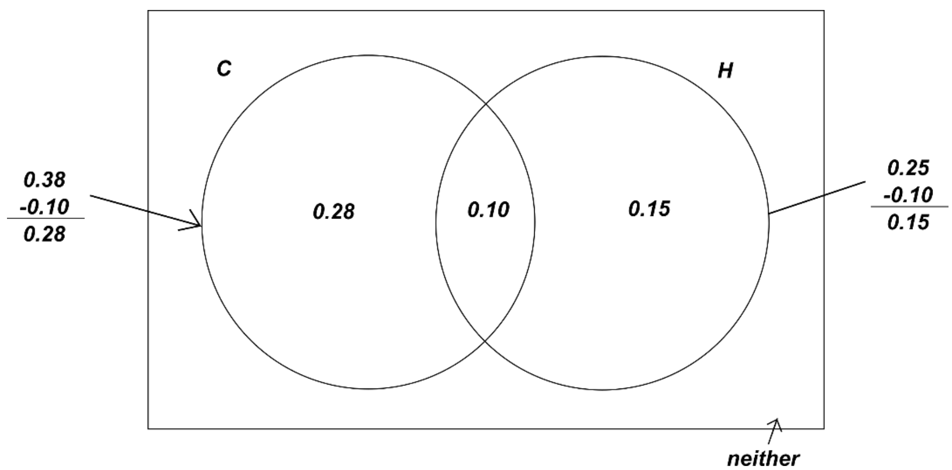
13. The answer is B). Pr(B/A) = $\frac{\Pr(A \text{ and } B)}{\Pr(A)} = \frac{1/8}{1/2} = \frac{1}{8} \times \frac{2}{1} = \frac{2}{8}$
= 0.25

14. The answer is A). $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{210-200}{10.2/\sqrt{5}} = 2.19$

$\Pr(z < 2.19) = 0.9857$

$\Pr(z > 2.19) = 1 - 0.9857$
= 0.0143

15.



The answer is D). $\Pr(C \text{ or } H) = \Pr(C) + \Pr(H) - \Pr(both)$
= 0.38 + 0.25 – 0.10
= 0.53

$\therefore \Pr(\text{like neither}) = 1 - 0.53 = 0.47$

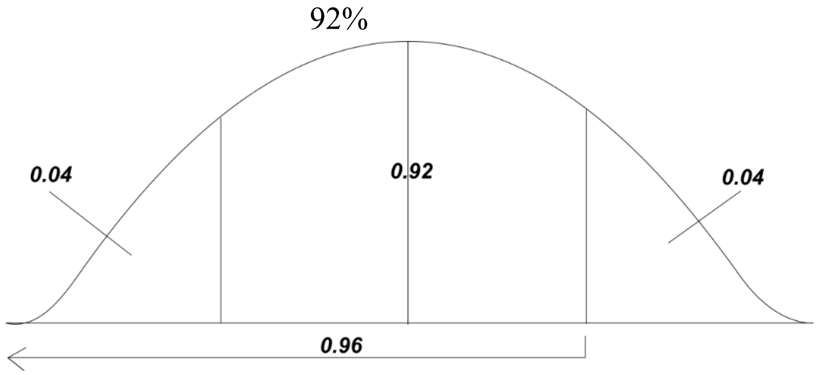
16. The answer is C). within 1 $\therefore m = 1$ $\sigma = 2.2$
90% CI $\therefore Z^* = 1.645$

$$n = \left(\frac{Z^*\sigma}{m}\right)^2 = \left(\frac{1.645(2.2)}{1}\right)^2 = 13.097 \quad \therefore 14 \text{ turkeys}$$

17. The answer is B). $\Pr(G/E^c) = \frac{\Pr(G \text{ and } E^c)}{\Pr(E^c)} = \frac{0.10}{1-0.40}$
$= \frac{0.10}{0.60} = 0.167$

18. The answer is C). Pr(BG) + Pr(GB)
$= 0.52\times 0.48 + 0.48 \times 0.52$
= 0.4992

19. The answer is D). $\bar{x} = 3.2 \quad n = 12 \quad \sigma = 0.2$



Z^* =look up 0.96 in body = 1.75

$$\mu = \bar{x} \pm Z^*\left(\frac{\sigma}{\sqrt{n}}\right) = 3.2 \pm 1.75\left(\frac{0.2}{\sqrt{12}}\right)$$
$$= 3.2 \pm 0.10$$
$$= (3.1, 3.3)$$

Total for 12 nails =(3.1(12), 3.3(12)) = (37.2, 39.6)

20. The answer is A). New margin of error is smaller by $\sqrt{9} = 3\ times$ since the sample size is 9 times larger, so new m= $\frac{120}{\sqrt{9}} = 40$

New m= old m $\div Z^* \times new\ Z^* = 40 \div 1.645 \times 1.96 = 47.7$

21. The answer is B).

Pr(A) = Pr(2, 3, 4, 5, 6 or more) = 0.9
Pr(B) = Pr(0, 1, 2, 3, 4) = 0.1 + 0.15 + 0.2 + 0.25 = 0.70
Pr(A and B) = Pr(2, 3, 4)
=0.15 + 0.2 + 0.25 = 0.60
Pr(A or B) = Pr(A)+Pr(B) – Pr(A and B)
= 0.9 + 0.7 – 0.6
= 1

22. The answer is C). Pr(B/A)= $\frac{\Pr(A\ and\ B)}{\Pr(A)} = \frac{0.60}{0.90} = 0.67$

23. The answer is D). $m = \frac{Z^*\sigma}{\sqrt{n}} = \frac{2.576(6)}{\sqrt{30}} = 2.82$

24. The answer is C). $m = Z^*\left(\frac{\sigma}{\sqrt{n}}\right) \qquad m = \frac{4.2-2.6}{2} = 0.8$

$$0.8 = 2.326\left(\frac{\sigma}{\sqrt{100}}\right)$$

$\sigma = 3.4$

25. The answer is B). $b = r\,\frac{s_y}{s_x}$

$-1.2 = r\,\frac{\sqrt{10}}{\sqrt{6}}$

-1.2 = 1.290994449 r

$r = -0.9295$

$r^2 = (-0.9295)^2 = 86\%$

26. The answer is D). $Z_1 = \frac{x_1-\mu}{\sigma} = \frac{8-10}{3.2} = -0.63$

$Z_2 = \frac{14.2-10}{3.2} = 1.31$

$\Pr(z < 1.31) - \Pr(z < -0.63)$

= 0.9049 – 0.2643 = 0.6406

27.

	60 - 69	70 - 79	80 - 89	90 +
Male	15	10	5	2
Female	20	15	10	x

The answer is A). Complete the table, find x

$\frac{20+15+10}{f} = 0.95$

$\frac{45}{f} = 0.95$

$f = 47$

∴ over 90+ = 47 – 20 – 15 – 10 = 2

Pr (F/80+)=$\frac{\frac{10+2}{47}}{\frac{5+10+2+2}{47}} = \frac{12}{19} = 0.63 \;\; or\ 63\%$

NOTE: the bottom is those 80+

28. The answer is B). $\bar{x} = \frac{100+150+225}{3} = 158.3$

$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\frac{(100-158.3)^2+(150-158.3)^2+(225-158.3)^2}{2}}$

$= \sqrt{\frac{3398.89 + 68.89 + 4448.89}{2}}$

$s = 62.9$

$SE = \frac{s}{\sqrt{n}}$

SE=$\frac{62.9}{\sqrt{3}} = 36.3$

29. The answer is C). Since 15 is in the interval, it is not statistically significant. Since 23 is in the interval, 23 is significant.

30. The answer is C). If $n\ doubles, m = \frac{Z^*\sigma}{\sqrt{n}}$

$$\therefore m\ is\ divided\ by\ \sqrt{2}$$
$$\therefore \frac{21.4}{\sqrt{2}} = 15.1$$

Long answer:

1. a) $\mu = \bar{x} \pm Z^*\left(\frac{\sigma}{\sqrt{n}}\right) = 65 \pm 1.96\left(\frac{5.3}{\sqrt{50}}\right) = 65 \pm 1.47$
   = (63.53, 66.47)

   b) $m = Z^*\left(\frac{\sigma}{\sqrt{n}}\right) = 2.576\left(\frac{5.3}{\sqrt{50}}\right) = 1.93$

   c) $n = (\frac{Z^*\sigma}{m})^2 = (\frac{1.645(5.3)}{1.93\times 2})^2 = 5.1 \quad \therefore n = 6$

2. $\bar{x} = \frac{14+19}{2} = 16.5$

   Margin of error = m = $\frac{19-14}{2} = 2.5$

   $new\ M = old\ M \div old\ Z^* \times new\ Z^*$
   $= 2.5 \div 2.576 \times 1.96$
   = 1.90
   $new\ CI = (\bar{x} - M, \bar{x} + M)$

   $(16.5 - 1.9, 16.5 + 1.9)$
   $= (14.6, 18.4)$

3.a) explanatory variable – red meat
   Response variable – cholesterol levels
  b) lurking variables – exercise
   - alcohol

4. a) two 4's – should be only one 4

b) 10, 21, 22, 31, 33, 34, [40], 41, 42, 51, 63, 71, 82
   median = 40
   mean = $\frac{10+21+\cdots+82}{13} = \frac{541}{13} = 41.6$

Q1 = (22+31)/2 = 26.5

c) right-skewed

5.

Exposed	Lung cancer	No lung cancer
Yes	20 a	85 b
No	5 c	100 d

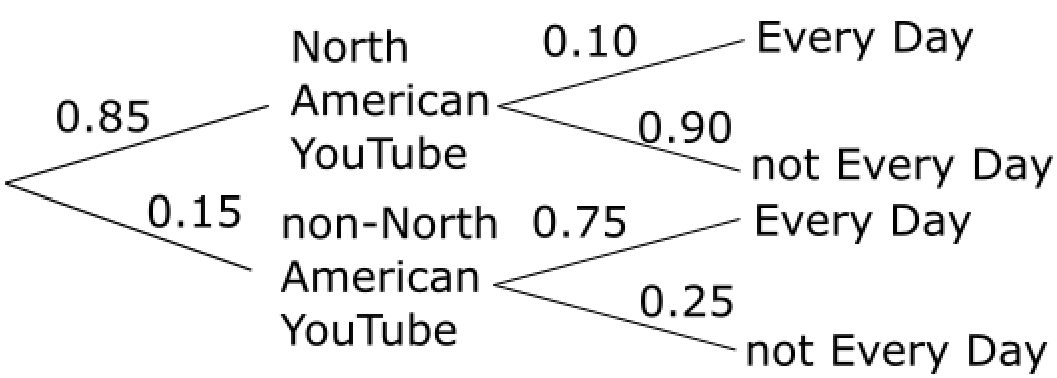
odds ratio $=\frac{ad}{bc} = \frac{20(100)}{85(5)} = 4.7$

$\therefore$ *The* group of smokers has approximately 5 times the odds of having lung cancer than non-smokers.

Relative risk $= \frac{a(c+d)}{c(a+b)} = \frac{20(5+100)}{5(20+85)} = 4$

$\therefore$ The relative risk for smokers developing lung cancer is 4 times that of non-smokers developing lung cancer.

6.a) tree diagram



	Use every day	Not every day
North American	0.10	0.90
Non-North American	0.75	0.25

b) $\Pr(notN/notE) = \frac{P(not\ N\ and\ not\ E)}{\Pr(not\ E)}$

$$= \frac{0.15(0.25)}{0.85(0.90) + 0.15(0.25)} = 0.047$$

7. $m = Z^*\left(\frac{\sigma}{\sqrt{n}}\right)$

$9.8 = Z^*(5)$

$Z^* = 1.96, so\ it\ is\ a\ 95\%\ confidence\ interval$

8.

Age Groups	Fail/Success	Treatment A	Treatment B	Total
< 40	Fail	5	35	40
	Success	80	235	315
40 +	Fail	70	25	95
	Success	190	50	240

Combined data:

Fail/Success	Treatment A	Treatment B	Total
Fail	75	60	135
Success	270	285	555

b) Calculate the success rates for treatments A and B when the data is split by age groups. Which treatment is better?

Treatment A

< 40 Success $= \frac{80}{85} = \boxed{0.941}$

$40 +$ Success $= \frac{190}{190+70} = \frac{190}{260} = \boxed{0.731}$

Treatment B

< 40 Success $= \frac{235}{35+235} = \frac{235}{270} = \boxed{0.870}$

$40 +$ Success $= \frac{50}{25+50} = \frac{50}{75} = \boxed{0.667}$

∴ the success rate is higher in both age groups for treatment *A* than treatment *B*

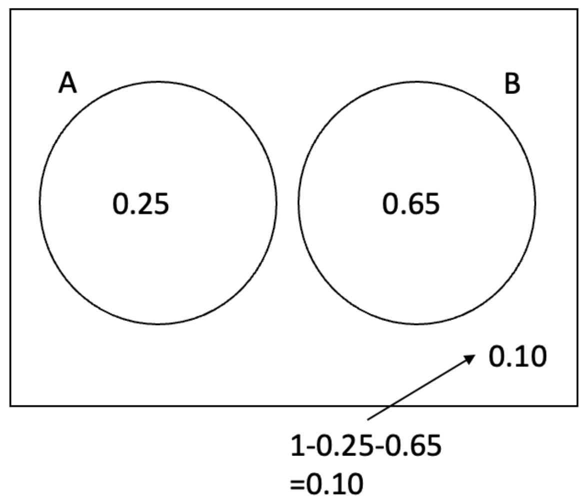
b) Calculate the success rates for treatments A and B when the data is combined. Which treatment has a higher success rate?

Combined	Treatment A	Treatment B
Success rate	$= \frac{270}{270+75} = \boxed{0.783}$	$= \frac{285}{60+285} = \frac{285}{345} = \boxed{0.826}$

∴ when we combine the data, treatment B has a higher success rate than treatment A

c) From a) and b) is this an example of Simpson's Paradox? Why or why not?

Yes, this is an example of Simpson's Paradox because when the data was separated by age groups, Treatment A had a higher success rate for each age group. However, once the data was combined, Treatment B has a higher success rate. When the relationship reverses when the data is combined, this is what is referred to as Simpson's Paradox.



9.a)
$= \Pr(A) + \Pr(B) - \Pr(A \cap B)$
$= 0.25 + 0.65 - 0$
$= 0.90$

$b) = \Pr(A) = 0.25$
Since all of A is outside of B

$c) \ = 0.10$
d) = Pr(B) = 0.65 since all of B is outside of A

10. Using Baye's Theorem:

$$\Pr(B/A) = \frac{\Pr(A/B)Pr(B)}{\Pr(A)} = \frac{\frac{1}{3}\left(\frac{1}{2}\right)}{\frac{1}{4}} = \frac{1}{6}\left(\frac{4}{1}\right) = \frac{4}{6} = \frac{2}{3}$$

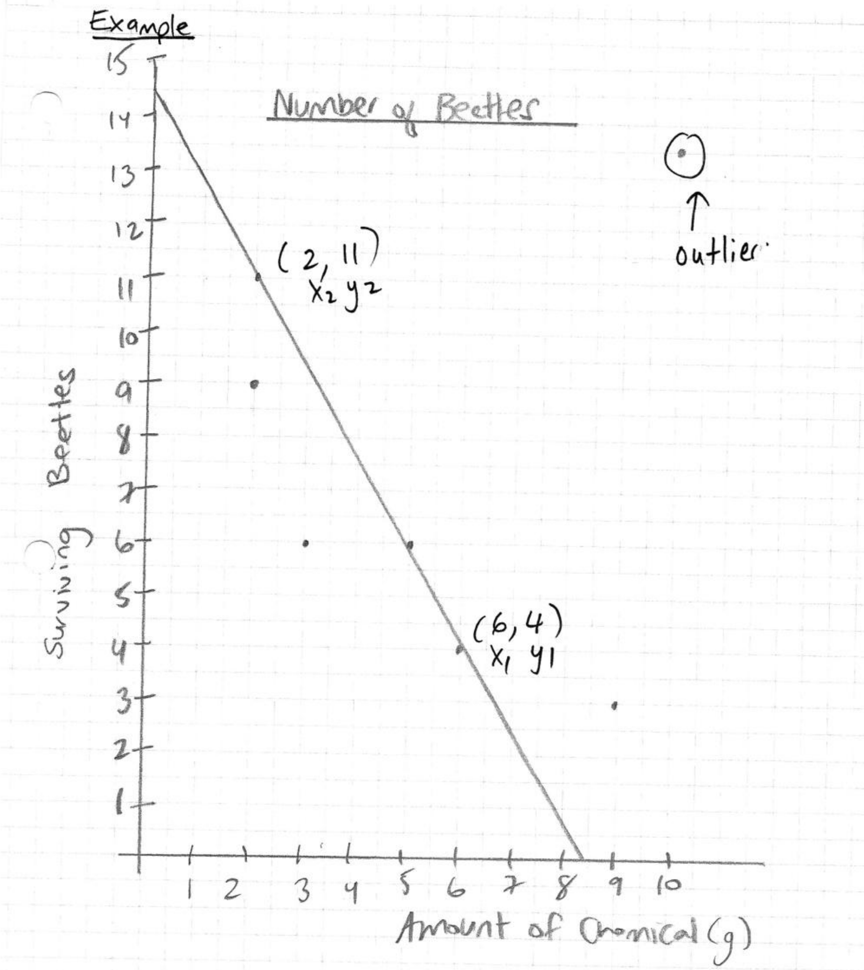
11.The explanatory variable is the amount of chemical being applied in grams. The response variable is the number of surviving beetles.

a) $b = \frac{y2-y1}{x2-x1} = \frac{11-4}{2-6} = -\frac{7}{4}$ (slope)

y-intercept is 14.5 (where the graph crosses the x axis)

Equation would be $\hat{y} = 14.5 - \frac{7}{4}x$

c) There is an outlier at (10,14).



d) 4.2 g (answers vary based on your graph)

e) $\hat{y} = 14.5 - \frac{7}{4}x$ let x=4 and solve for y

$$\hat{y} = 14.5 - \frac{7}{4}x(4)$$
$$\hat{y} = 7$$

So, there would be 7 surviving beetles.

12.

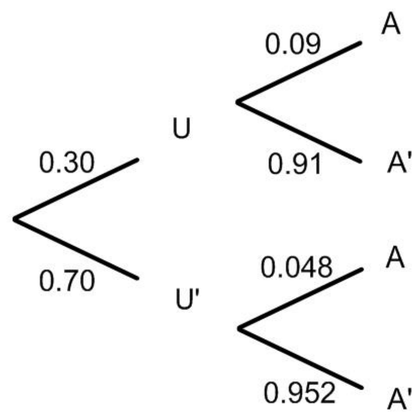
$U = under\ 25$

$P(U) = 0.30\ \ under\ 25$

$$\Pr(U/A) = \frac{\Pr\ (U\ and\ A)}{\Pr\ (A)}$$

$$= \frac{0.3\times0.09}{0.3\times0.09+0.7\times0.048}$$

$$= 0.446$$



13.

a) We are 95% confident the population mean is between 10 and 30.

b) If we suppose the mean, $\mu = 32$, then since 32 is NOT in between 10 and 30, we would reject that claim and say there is statistically significant evidence that the mean is NOT equal to 32.

c) **Bootstrap** is a type of resampling in which we can make inferences about a population from which our data is a random sample. We can use these methods to make inferences about other parameters, besides the mean.

In this method we generate many samples by sampling with replacement from your original sample. These samples are referred to as bootstrap samples.

d) $\mu = \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$ is the form of a confidence interval

The sample mean is $\bar{x}$ and the margin of error is m=$z^* \frac{\sigma}{\sqrt{n}}$ and the critical number is z^*.

Best of luck

on your exam!!!!!!